

ADR

E-FILING

ORIGINAL FILED

MAR 31 2004

Richard W. Wiekling  
Clark, U.S. District Court  
Northern District of California  
San Jose

MILBANK, TWEED, HADLEY & McCLOY LLP  
James Pooley (CA Bar No. 058041)  
L. Scott Oliver (CA Bar No. 174824)  
Marc David Peters (CA Bar No. 211725)  
Anupam Sharma (CA Bar No. 229545)  
3000 El Camino Real  
Five Palo Alto Square, 7<sup>th</sup> Floor  
Palo Alto, California 94306-2109  
Telephone: (650) 739-7000  
Facsimile: (650) 739-7100

Attorneys for Plaintiff UniRAM TECHNOLOGY, INC.

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

UniRAM TECHNOLOGY, INC., a  
California corporation,

Plaintiff,

v.

MONOLITHIC SYSTEM  
TECHNOLOGY, a Delaware  
corporation,

Defendant.

CASE NO.

004 01268 *PK*

COMPLAINT FOR TRADE SECRET  
MISAPPROPRIATION, PATENT  
INFRINGEMENT, INTERFERENCE  
WITH CONTRACT, AND UNFAIR  
COMPETITION; DEMAND FOR JURY  
TRIAL

# PARTIES

1. Plaintiff UniRAM Technology, Inc. ("UniRAM") is a corporation organized under the laws of California and has a principal place of business at 3375 Scott Boulevard, Santa Clara, California.

2. Defendant Monolithic System Technology, Inc. ("MoSys") is a corporation organized under the laws of Delaware and has a principal place of business at 1020 Stewart Drive, Sunnyvale, California.

///

///

///

PA1:#24007975v9

COMPLAINT FOR TRADE SECRET MISAPPROPRIATION,  
PATENT INFRINGEMENT, ETC. / DEMAND FOR JURY  
TRIAL

MILBANK, TWEED, HADLEY & McCLOY LLP  
ATTORNEYS AT LAW  
PALO ALTO

COPY

## JURISDICTION AND VENUE

3. This is an action for damages and injunctive relief based upon patent infringement arising under Title 35 of the United States Code, upon trade secret misappropriation under California Civil Code section 3426, upon interference with contract, and upon violations of California unfair competition law.

4. This Court has jurisdiction over the subject matter of this action pursuant to 28 U.S.C. § 1331 and 28 U.S.C. § 1338. Supplemental subject matter jurisdiction for the state law claims is based on 28 U.S.C. § 1367. Venue is proper in this District pursuant to 28 U.S.C. § 1391.

5. This Court has jurisdiction over MoSys because MoSys has committed, induced, and/or contributed to acts of patent infringement, misappropriated trade secrets, and committed acts of unfair competition during the course of its business in this District.

## THE PATENTS IN SUIT

6. On August 22, 2000, U.S. Patent No. 6,108,229 ("the '229 patent"), entitled "High Performance Embedded Semiconductor Memory Device With Multiple Dimension First-Level Bit-Lines" was duly and legally issued to Jeng-Jye Shau. UniRAM is the owner of the entire right, title, and interest in and to the '229 patent. A copy of the '229 patent is attached as Exhibit A.

7. On February 3, 2004, U.S. Patent No. 6,687,148 B2 ("the '148 patent"), entitled "High Performance Embedded Semiconductor Memory Device With Multiple Dimension First-Level Bit-Lines" was duly and legally issued to Jeng-Jye Shau. UniRAM is the owner of the entire right, title, and interest in and to the '148 patent. A copy of the '148 patent is attached as Exhibit B.

## THE HIGH-PERFORMANCE COMPUTER MEMORY PROBLEM

8. In the mid-1990s, Dr. Shau, the inventor of the '229 and '148 patents, realized that the development of semiconductor memory circuits was falling behind the pace of improvements in logic integrated circuits (commonly referred to as "processors" and in some cases as "CPUs"). In short, he anticipated that by the late 1990s, memories would be the limiting factor for

1 computer evolution. Dr. Shau, who had formerly worked with Intel on microprocessor design,  
2 founded his own company (the predecessor in interest to UniRAM) to set about solving this  
3 problem.

4 9. Conventional computer memory circuits are of two basic types. The first, known as  
5 “DRAM,” combines one transistor and one capacitor. The capacitor can store an electrical  
6 charge for a short time, and the presence or absence of that electrical charge represents one bit of  
7 information—a zero or a one. DRAM circuits can be dense (i.e., many such transistor/capacitor  
8 sets can be packed onto a single chip), but they have certain drawbacks. First, the speed at which  
9 information can be accessed is usually slow. Second, and related, the capacitors storing the  
10 electrical charges will leak over time, so the information in the memory circuit needs to be  
11 periodically refreshed, just as a bucket of water with a hole in the bottom will store water for  
12 some period before it needs to be topped off again. And during that topping off process, no  
13 information can be retrieved from that DRAM circuit, forcing the processor to wait.

14 10. The second conventional type of memory circuit is known as “SRAM.” Rather than  
15 using a transistor and capacitor, an SRAM circuit uses six transistors (or in some cases four).  
16 Those transistors form a circuit that can be toggled back and forth between two states, like a  
17 household light switch. One state represents a zero; the other, a one. SRAM circuits are faster  
18 than DRAM, and because there is no gradual draining of an electrical charge, they do not need to  
19 stop and top off their stored data. The chief problem with SRAM is that it consumes a relatively  
20 large amount of space on a computer chip: instead of taking up only enough room for a single  
21 transistor and capacitor, it requires space and interconnection among six transistors. Space on a  
22 computer chip is always a precious commodity.

23 11. Also in the mid-1990s, and continuing today, chip designers sought to combine logic  
24 and memory onto a single chip. In a mobile telephone, for instance, having separate processor  
25 and memory chips takes up valuable space and costs more to build than a mobile telephone that  
26 has its processor and its memory combined on the same chip. “System-On-a-Chip”  
27 combinations provide tremendous advantages to designers, but the actual production of memory  
28

1 and logic (processor) circuits on the same chip has been extremely difficult. Memory and logic  
2 circuits tend to be built with different—and incompatible—processes. Moreover, even if it were  
3 possible to combine the two processes, doing so would add many extra steps to the fabrication,  
4 and every step increases the likelihood that a defect will creep into the circuit being built.

5 12. In short, then, computer memory needed to be fast, it needed to be small, and its  
6 construction and operation needed to be compatible with nearby logic circuits on the same chip.  
7 In the mid-1990s, these problems were daunting. Dr. Shau's solution was a radically new  
8 approach to building memory circuits.

### 9 DR. SHAU'S SOLUTION

10 13. Dr. Shau's inspiration was to combine the small size of DRAM with the speed of  
11 SRAM, and—crucially—to build these new memory circuits *during* the logic circuit fabrication,  
12 eliminating the problems of incompatible processes and many extra steps.

13 14. The solution is simple in principle; in execution, it is complex. Dr. Shau developed a  
14 solution, and on May 24, 1996, he filed for the first in a series of patents on this technology.  
15 During the remainder of 1996—and throughout 1997 and 1998—Dr. Shau and his company  
16 invested considerable sums of money and engineering time proving that his technical solutions  
17 could be efficiently produced on a large scale.

### 18 DR. SHAU'S NON-DISCLOSURE AGREEMENTS WITH TSMC

19 15. Because only a handful of very large companies have semiconductor fabrication  
20 plants (often called “fabs”) of their own, most semiconductor design firms arrange to have their  
21 chip designs built by specialized fab companies. One of the largest and oldest of such contract  
22 fabs is Taiwan Semiconductor Manufacturing Co., Ltd., or “TSMC.” TSMC works with  
23 companies around the world, building chips from those companies' designs. TSMC's work is  
24 widely recognized as high-caliber, and Dr. Shau was interested in building his chips at TSMC.

25 16. Since TSMC is constantly exposed to innovative designs from its customers, many of  
26 whom are in competition with each other, it is accustomed to signing and abiding by non-  
27 disclosure agreements with its customers to ensure that their valuable trade secrets are kept



1 confidential. Moreover, because TSMC is only interested in working with a small company if it  
2 can be convinced that the company's technology is successful and will generate substantial sales  
3 for TSMC, TSMC demanded full disclosure of UniRAM's technology under an NDA.

4 17. On September 16, 1996, Dr. Shau's company at the time (Telesis Innovation, Inc.—a  
5 predecessor in interest to UniRAM) and TSMC executed an NDA under which TSMC agreed to  
6 protect the secrecy of Dr. Shau's invention. A copy of that NDA is attached as Exhibit C. They  
7 executed additional NDAs on October 11, 1999 (between InTempo—another predecessor in  
8 interest to UniRAM—and TSMC, attached as Exhibit D), and August 29, 2000 (between  
9 UniRAM and TSMC, attached as Exhibit E). TSMC and UniRAM remain subject to an NDA  
10 today.

#### 11 **DR. SHAU'S DISCLOSURES OF TRADE SECRETS TO TSMC**

12 18. Once an NDA was in place, Dr. Shau revealed his inventions to TSMC so that TSMC  
13 would understand the value of his inventions and agree to build circuits for UniRAM. The first  
14 presentation was late 1996 to TSMC sales managers, most of whom had Ph.D. degrees in  
15 engineering. The audience for this first presentation was sufficiently impressed that the current  
16 president of TSMC, Mr. F.C. Tseng, telephoned Dr. Shau to praise the invention and express  
17 TSMC's interest.

18 19. TSMC asked Dr. Shau to make a series of detailed presentations to TSMC in early  
19 1997. Attendees from TSMC included the Vice President for Corporate Research and  
20 Development, the Deputy Director of the Embedded DRAM Product Management Program, the  
21 Program Deputy Director of the Memory Technology Development Division for Research and  
22 Development, and the Deputy Director of Technical Marketing for the Corporate Marketing  
23 Division.

24 20. In the presentations, Dr. Shau explained how to build embedded memory devices  
25 using a logic fabrication process. This was a revolutionary advance, and TSMC expressed great  
26 interest in the value of Dr. Shau's technology. Dr. Shau told TSMC that he expected to serve a  
27 significant portion of the market for logic devices made with embedded memory.

1           21. Dr. Shau explained to TSMC that its then-existing embedded DRAM devices were  
2 difficult to construct because of conflicts between the memory and logic processes. His advance  
3 was to build embedded DRAM with a logic, or near-logic, process, but implementing that  
4 advance would require that Dr. Shau work closely with TSMC. Dr. Shau also told TSMC that  
5 his architecture was to break up large memory arrays into very small blocks, creating a high  
6 signal-to-noise ration, and to use error correction circuitry to ensure reliability.

7           22. At the same time, TSMC and Dr. Shau discussed plans for rolling out Dr. Shau's  
8 second and third generation devices. The first generation approach—building a memory circuit  
9 with a logic circuit fabrication process (known as "1T," for "1 transistor")—provided very good  
10 results once the fabrication techniques were ironed out. During the development stage, Dr. Shau  
11 made detailed studies of the problems, and worked with TSMC manufacturing data to develop  
12 solutions to problems encountered in development. Some of Dr. Shau's solutions to the  
13 problems were the subject of continuation patent applications. Dr. Shau's 1T circuits had very  
14 high speed, low power consumption, and good space savings over SRAM 6 or 4 transistor  
15 memory technology. Dr. Shau achieved part of his success by using small blocks of memory  
16 cells that could react to information requests or be refreshed at very high speed. The first  
17 generation process also included error correction circuitry to detect any anomalies in the data  
18 retrieved from the memory cells, and it required no additional steps in fabrication. But Dr. Shau  
19 had already developed innovative improvements which he also disclosed in confidence to  
20 TSMC.

21           23. Dr. Shau also developed a variation of the first generation device, planned for testing  
22 in 1997, which requires one more step in the fabrication process, but increases the density of the  
23 memory cells by twenty percent. Given that space on a chip is always at a premium, this savings  
24 was well worth the extra step.

25           24. The second generation device involved two extra steps during fabrication. In those  
26 steps, the first generation 1T process was modified by placing a "shallow trench" near the  
27 memory cell's transistor. In essence, this trench would create a vertical capacitor, rather than a  
28

1 horizontal one, thus occupying much less space on the chip. While conventional DRAM circuits  
2 can use similar vertical capacitors, one of Dr. Shau's innovations was to build the "shallow  
3 trench" using a logic fabrication process. The resulting capacitor had more leakage and a smaller  
4 stored charge than a typical DRAM capacitor, but Dr. Shau also developed methods of  
5 overcoming those problems. The second generation device can reach near-DRAM memory cell  
6 densities, while retaining extremely high speeds that be unattainable with DRAM technology—  
7 all without causing a significant increase in the manufacturing cost.

8 25. The third generation device used conventional DRAM storage capacitors, coupled  
9 with a logic transistor acting as a "word line" transistor in the memory cell. This process  
10 achieves the best density—allowing more memory cells to be packed onto a single chip—while  
11 reducing manufacturing complexities (typical DRAM transistors are high voltage, "thick gate"  
12 devices that are difficult to build using standard logic processes.) In addition, Dr. Shau further  
13 increased the speed of this third generation device by using a small memory block design.

14 26. In their meetings, Dr. Shau disclosed to TSMC all details, including memory cell and  
15 circuit design key factors and architecture, and the parties agreed this information was trade  
16 secret and was protected by the parties' NDA. Dr. Shau explained to TSMC that he was seeking  
17 a long-term relationship with a reliable fab so both parties could reap the rewards of his  
18 inventions.

19 27. Dr. Shau also told TSMC in confidence that his goal was to begin 1T production in  
20 1997, and testing the third generation device by early 1998.

21 28. In a recap of these plans that Dr. Shau presented to TSMC in early 1997, it was  
22 acknowledged that TSMC would only be providing "foundry" (chip fabrication) services, and  
23 that no licensing or partnership with UniRAM was contemplated. The value of the relationship  
24 for TSMC would come through its fabrication fees, while UniRAM would realize on the value of  
25 its innovations through sales of chips and licenses to make them.

26 29. In the second half of 1998, UniRAM's predecessor in interest completed the design  
27 for a 128k Cache RAM circuit using UniRAM's technology. Strangely, TSMC held that design  
28

1 for nearly six months, finally building parts from it in April, 1999, and collecting its fabrication  
2 fees. However, unknown to UniRAM at the time, a competitor was about to emerge, a company  
3 that had improperly gained access to its secret technology.

#### 4 **MOSYS ACQUIRES UNIRAM'S TRADE SECRETS**

5 30. MoSys began working with TSMC in the mid-1990s, and to help pay for the parts  
6 TSMC fabricated, MoSys transferred 5% of its stock to TSMC in late 1996. After MoSys' IPO  
7 in 2001, TSMC continued to own 2% of MoSys' stock.

8 31. In the mid-1990s, MoSys developed proprietary memory banks that could operate at  
9 reasonably high speed—an attractive feature for high-resolution computer graphics components,  
10 which need the ability to store and display graphics (for instance, for use in video games)  
11 quickly. However, MoSys' proprietary technology prevented mainstream adoption, and its sales  
12 were weak.

13 32. In 1996, MoSys released a new DRAM-based memory chip that provided reasonable  
14 speed, was somewhat smaller and cost less than comparable SRAM technology. MoSys  
15 continued to use non-standard interfaces for these parts, which limited consumer acceptance, and  
16 oversupply in the SRAM market led to a collapse in SRAM prices, erasing MoSys' price  
17 advantage. Sales, again, were unimpressive.

18 33. MoSys went back to what it termed "research and development" for nearly two years.  
19 Prior to 1998, all of MoSys' products were stand-alone memory circuits using conventional  
20 DRAM designs; before 1998, MoSys had never built an embedded ("System on a Chip")  
21 memory circuit. But in 1998, MoSys released a 1T memory cell built using a logic fabrication  
22 process. Its technology bears an uncanny resemblance to UniRAM's inventions.

23 34. Like UniRAM, MoSys' 1T memories are made using standard logic fabrication  
24 methods. Like UniRAM, MoSys' 1T memories are extremely fast because MoSys used small  
25 blocks of memory cells. And like UniRAM, MoSys includes error correction circuitry to ensure  
26 the reliability of data retrieved from the memory blocks.



35. MoSys has even announced second (1T-SRAM-M), third (1T-SRAM-R), and fourth (1T-SRAM-Q) generation memory technologies that tracked UniRAM's secret roadmap. Indeed, MoSys has recently begun licensing a vertical capacitor design—1T-SRAM-Q—the key to UniRAM's second generation device. Since 1998, all of the significant technology “developments” made by MoSys followed, step by step, the road maps UniRAM had earlier provided to TSMC.

36. That MoSys' business shifted to mimic UniRAM's cannot be explained by coincidence, nor can the striking similarity of its technology be squared with independent development. UniRAM is informed and believes, and thereon alleges, that MoSys acquired UniRAM's trade secrets from TSMC, and that MoSys then set about deliberately copying UniRAM's inventions.

#### **MOSYS' PROFITS FROM ITS USE OF UNIRAM'S TRADE SECRETS**

37. MoSys' plan seems to have brought it near-term success. Its first major design win came in 1999, when it licensed embedded memory to Nintendo for use in Nintendo's forthcoming video game console. It followed that transaction by licensing similar technology to companies such as Sanyo and Fujitsu. Significantly, it also granted a fabrication license for that technology to TSMC.

38. In 2002, MoSys realized revenue for the sales of products totaling nearly \$3 million, almost \$11 million in license fees, and more than \$14 million in royalties. In short, MoSys earned \$28 million by using technology misappropriated from UniRAM.

39. In 2003, MoSys' sales totaled close to \$20 million, and its plans to release a fourth-generation product (1T-SRAM-Q) using UniRAM's technology are touted by MoSys as driving its expected 2004 and 2005 revenues. In the space of two years, MoSys' misappropriation of UniRAM's trade secrets has yielded it almost \$50 million.

#### **MOSYS' PAST AND FUTURE PATENT INFRINGEMENT**

40. UniRAM filed its first patent application on its memory cell inventions in 1996. By the time its trade secrets were leaked to MoSys, UniRAM had 4 applications on file. At present,

1 UniRAM has a portfolio of more than a dozen issued patents on its first, second, and third  
2 generation memory technology. MoSys' future products—which UniRAM is informed and  
3 believes it plans to license to TSMC—will apparently be copies or derivatives of UniRAM's  
4 second generation device using the vertical capacitor, a feature patented by UniRAM since early  
5 1997.

#### 6 CLAIM I—TRADE SECRET MISAPPROPRIATION

7 41. UniRAM realleges and incorporates by reference each of the preceding paragraphs of  
8 this complaint.

9 42. UniRAM enjoys an advantage over its existing and would-be competitors in the  
10 design, development, production, promotion and marketing of advanced memory circuits  
11 because of the UniRAM trade secrets described above, including but not limited to its first,  
12 second, and third generation devices, its small blocks of memory cells, and its error correction  
13 circuitry.

14 43. UniRAM has made reasonable efforts under the circumstances to protect the  
15 confidentiality of its trade secrets, including the expression of some of those secrets in patent  
16 applications before they were published. UniRAM's confidential information shared with  
17 TSMC derives independent economic value from not being generally known to the public or  
18 other persons who could obtain economic value from its disclosure or use. Accordingly, this  
19 information qualifies as trade secrets under California's Uniform Trade Secrets Act, Cal. Civ.  
20 Code § 3426 *et seq.*

21 44. TSMC and its employees were under a duty to keep UniRAM's confidential  
22 information secret. MoSys knew or reasonably should have known that TSMC owed a duty to  
23 UniRAM to maintain the information in secrecy. Nevertheless, on information and belief,  
24 MoSys obtained this information through improper means and without the express or implied  
25 consent of UniRAM, and MoSys is now using the trade secrets in connection with its own  
26 business activities.

1           45. Each of the acts of misappropriation was, on information and belief, done willfully  
2 and maliciously by MoSys.

3           46. As a direct result of MoSys' misappropriation of UniRAM's trade secrets, MoSys has  
4 been unjustly enriched and UniRAM has sustained damages in an amount to be proven at trial.  
5 UniRAM has also suffered irreparable harm as a result of MoSys' activities and will continue to  
6 suffer irreparable injury that cannot be adequately remedied at law unless MoSys, its officers,  
7 agents, employees, and all persons acting in concert with it, are temporarily, preliminarily, and  
8 permanently enjoined from engaging in further such acts of misappropriation or enjoying the  
9 fruits of its misappropriation.

#### 10                           **CLAIM II—INFRINGEMENT OF THE '229 PATENT**

11           47. UniRAM realleges and incorporates by reference each of the preceding paragraphs of  
12 this complaint.

13           48. MoSys has infringed and continues to infringe; has induced and continues to induce  
14 others to infringe; and/or has committed and continues to commit acts of contributory  
15 infringement of one or more of the claims of the '229 patent. MoSys' infringing activities in the  
16 United States and this District include development, manufacture, use, sale, and/or offer for sale  
17 of UniRAM's memory cell technology. Such infringing activities violate 35 U.S.C. § 271.  
18 Upon information and belief, such infringement has been, and continues to be, willful.

19           49. As a consequence of the infringing activities of MoSys regarding the '229 patent as  
20 complained of herein, UniRAM has suffered monetary damages in an amount not yet  
21 determined, and UniRAM will continue to suffer irreparable damages in the future unless and  
22 until MoSys' infringing activities are enjoined by this Court.

#### 23                           **CLAIM III—INFRINGEMENT OF THE '148 PATENT**

24           50. UniRAM realleges and incorporates by reference each of the preceding paragraphs of  
25 this complaint.

26           51. MoSys has infringed and continues to infringe; has induced and continues to induce  
27 others to infringe; and/or has committed and continues to commit acts of contributory

1 infringement of, one or more of the claims of the '148 patent. MoSys' infringing activities in the  
2 United States and this District include development, manufacture, use, sale, and/or offer for sale  
3 of UniRAM's memory cell technology. Such infringing activities violate 35 U.S.C. § 271.  
4 Upon information and belief, such infringement has been, and continues to be, willful.

5 52. As a consequence of the infringing activities of MoSys regarding the '148 patent as  
6 complained of herein, UniRAM has suffered monetary damages in an amount not yet  
7 determined, and UniRAM will continue to suffer irreparable damages in the future unless and  
8 until MoSys' infringing activities are enjoined by this Court.

#### 9 CLAIM IV—INTENTIONAL INTERFERENCE WITH CONTRACT

10 53. UniRAM realleges and incorporates by reference each of the preceding paragraphs of  
11 this complaint.

12 54. The NDA agreements between TSMC and UniRAM (including UniRAM's  
13 predecessors in interest) are valid and enforceable contracts obligating TSMC to maintain  
14 UniRAM's trade secrets in confidence.

15 55. On information and belief, MoSys is and has been aware of the existence of these  
16 contracts between TSMC and UniRAM.

17 56. MoSys interfered with UniRAM's contractual NDAs with TSMC by obtaining and  
18 misusing the information UniRAM had disclosed in confidence to TSMC.

19 57. This interference has caused UniRAM to suffer monetary damages in an amount not  
20 yet determined, and it will continue to suffer irreparable damages in the future unless and until  
21 MoSys' interference is enjoined by this Court.

#### 22 CLAIM V—UNFAIR COMPETITION IN VIOLATION OF CALIFORNIA LAW

23 58. UniRAM realleges and incorporates by reference each of the preceding paragraphs of  
24 this complaint.

25 59. MoSys' actions complained of above constitute California common law unfair  
26 competition and violations of California Business and Professions Code § 17200 *et seq.*

27

28



**PRAYER FOR RELIEF**

WHEREFORE, Plaintiff UniRAM prays for judgment and relief as follows:

1. A declaration that MoSys has misappropriated UniRAM's trade secrets;
2. Temporary, preliminary, and permanent injunctions restraining MoSys, its officers, agents, servants, employees, attorneys, parents, subsidiaries, and other persons in concert or participation with it, from directly or indirectly obtaining, using, or communicating to any person or entity any trade secrets or confidential information of UniRAM;
3. An award of compensatory damages for trade secret misappropriation, including an accounting and award of all MoSys' gains, profits, and savings derived from its improper conduct;
4. An award of treble damages pursuant to Cal. Civil Code § 3426.3;
5. A declaration that MoSys has infringed, induced infringement of, and contributorily infringed, the '229 and '148 patents in violation of 35 U.S.C. § 271;
6. An award of actual and consequential damages pursuant to, *inter alia*, 35 U.S.C. § 284 in an amount appropriate to compensate UniRAM for the damages caused by MoSys' infringement of the '229 and '148 patents;
7. Temporary, preliminary, and permanent injunctions pursuant to 35 U.S.C. § 283 restraining MoSys, its officers, agents, servants, employees, attorneys, parents, subsidiaries, and other persons in concert or participation with it, from further infringing the '229 and '148 patents, and from making, using, offering for sale, licensing, importing or selling its infringing memory products;
8. A declaration that MoSys' infringement of the '229 and '148 patents has been willful and deliberate and that this case is exceptional pursuant to 35 U.S.C. §§ 284 and 285;
9. An award of treble damages pursuant to 35 U.S.C. § 284;
10. A declaration that MoSys has interfered with contracts between UniRAM and TSMC;
11. Temporary, preliminary, and permanent injunctions restraining MoSys, its officers, agents, servants, employees, attorneys, parents, subsidiaries, and other persons in concert or

1 participation with it, from directly or indirectly interfering with contracts between UniRAM and  
2 TSMC;

3 12. An award of compensatory damages for interference with contract, including an  
4 accounting and award of all MoSys' gains, profits, and savings derived from its improper  
5 conduct;

6 13. Damages in an amount to be determined at trial;

7 14. Restitution and disgorgement of defendant's unjust enrichment, including but not  
8 limited to all licensing revenues, royalties, and other sums earned by selling and/or licensing  
9 MoSys' 1T technology, pursuant to Cal. Bus. & Prof. Code § 17203;

10 15. An award of costs and attorneys' fees as permitted by law, including pursuant to 35  
11 U.S.C. § 285 and Cal. Civil Code § 3426.4;

12 16. An award of pre-judgment interest; and

13 17. Such other and further relief as the Court may deem just and proper.

14  
15 Dated: March 31, 2004

Respectfully submitted,

16 MILBANK, TWEED, HADLEY & McCLOY LLP

17  
18 By 

19 James Pooley

20 L. Scott Oliver

Marc David Peters

21 Anupam Sharma

22 Attorneys for Plaintiff UniRAM Technology, Inc.

**JURY DEMAND**

UniRAM demands a jury trial on all issues triable to a jury in this matter.

Dated: March 31, 2004

Respectfully submitted,

MILBANK, TWEED, HADLEY & McCLOY LLP

By: 

James Pooley

L. Scott Oliver

Marc David Peters

Anupam Sharma

Attorneys for Plaintiff UniRAM Technology, Inc.

MILBANK, TWEED, HADLEY & McCLOY LLP  
ATTORNEYS AT LAW  
PALO ALTO

# EXHIBIT A



**United States Patent** [19][11] **Patent Number:** **6,108,229****Shau**[45] **Date of Patent:** **Aug. 22, 2000**[54] **HIGH PERFORMANCE EMBEDDED SEMICONDUCTOR MEMORY DEVICE WITH MULTIPLE DIMENSION FIRST-LEVEL BIT-LINES**[76] Inventor: **Jeng-Jye Shau**, 991 Amarillo Ave., Palo Alto, Calif. 94303[21] Appl. No.: **09/114,538**[22] Filed: **Jul. 13, 1998****Related U.S. Application Data**

[63] Continuation-in-part of application No. 08/653,620, May 24, 1996, Pat. No. 5,748,547, and a continuation-in-part of application No. 08/805,290, Feb. 25, 1997, Pat. No. 5,825,704.

[51] **Int. Cl.<sup>7</sup>** ..... **G11C 5/02**[52] **U.S. Cl.** ..... **365/52; 365/104; 365/184**[58] **Field of Search** ..... **365/52, 72, 104, 365/102, 103, 184**[56] **References Cited****U.S. PATENT DOCUMENTS**

4,980,799	12/1990	Tobita	361/311
5,532,181	7/1996	Takebuchi et al.	437/43
5,912,489	6/1999	Chen et al.	257/321

*Primary Examiner*—David Nelms*Assistant Examiner*—Thong Le*Attorney, Agent, or Firm*—Bo-In Lin[57] **ABSTRACT**

A dynamic random access memory solves long-existing tight pitch layout problems using a multiple-dimensional bit line structure. Improvement in decoder design further reduces total area of this memory. A novel memory access procedure provides the capability to make internal memory refresh completely invisible to external users. By use of such memory architecture, higher performance DRAM can be realized without degrading memory density. The requirements for system support are also simplified significantly.

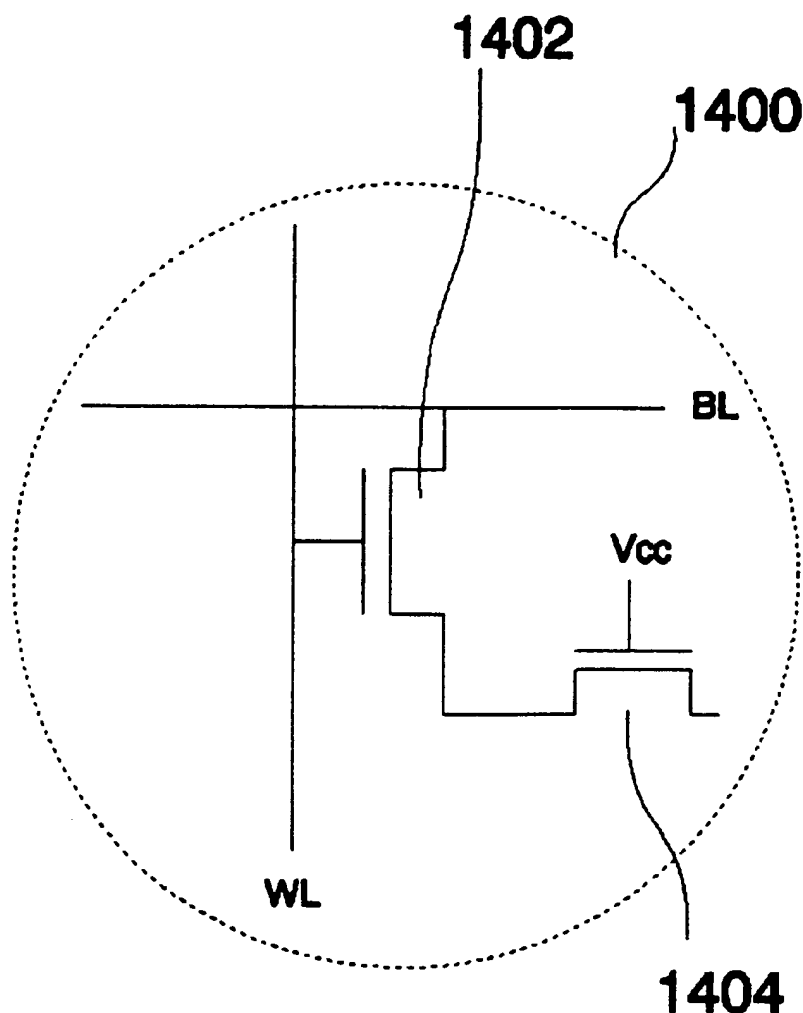
**5 Claims, 30 Drawing Sheets**

FIG.1 (Prior Art)

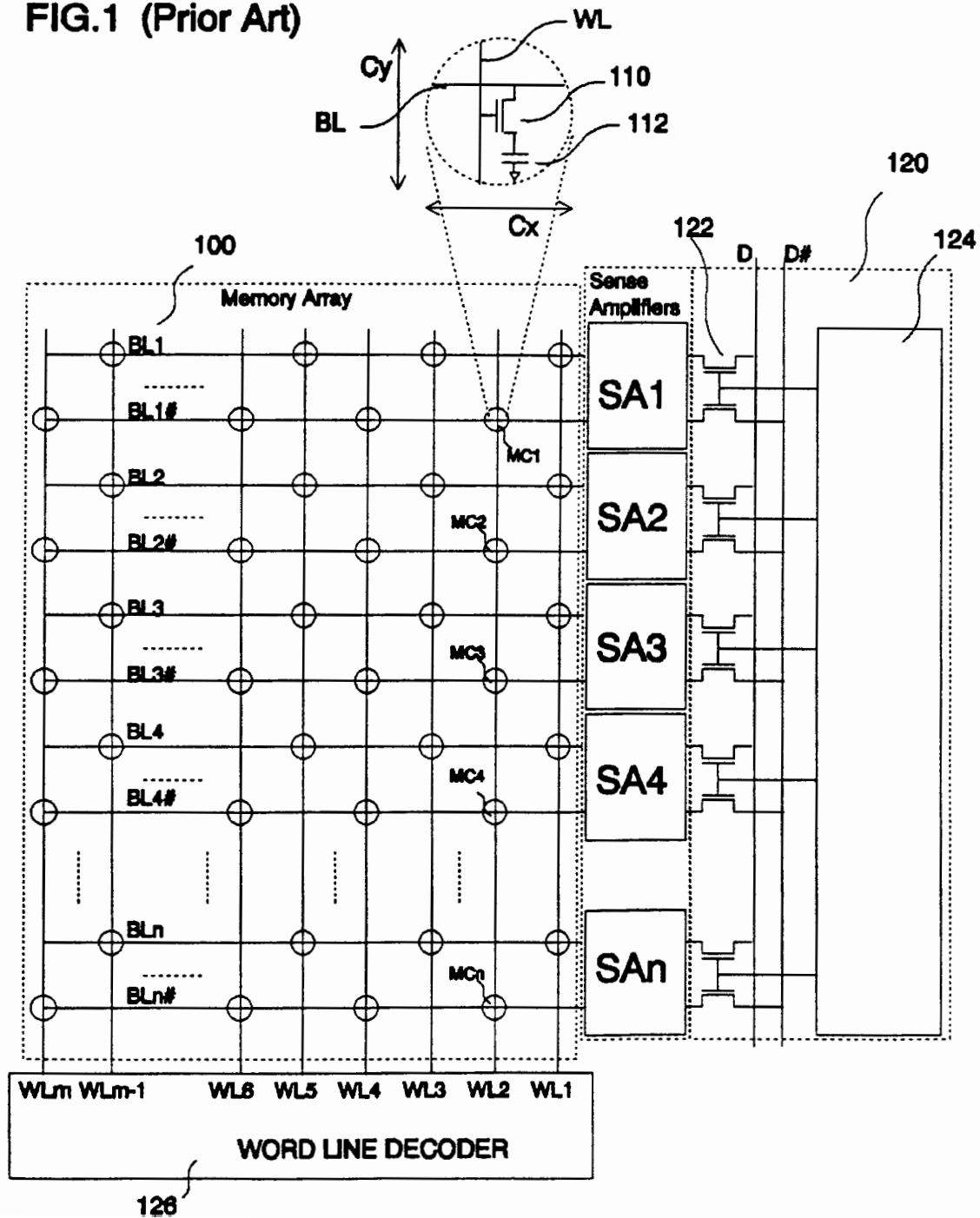


FIG. 2 (prior art multi-bank memory)

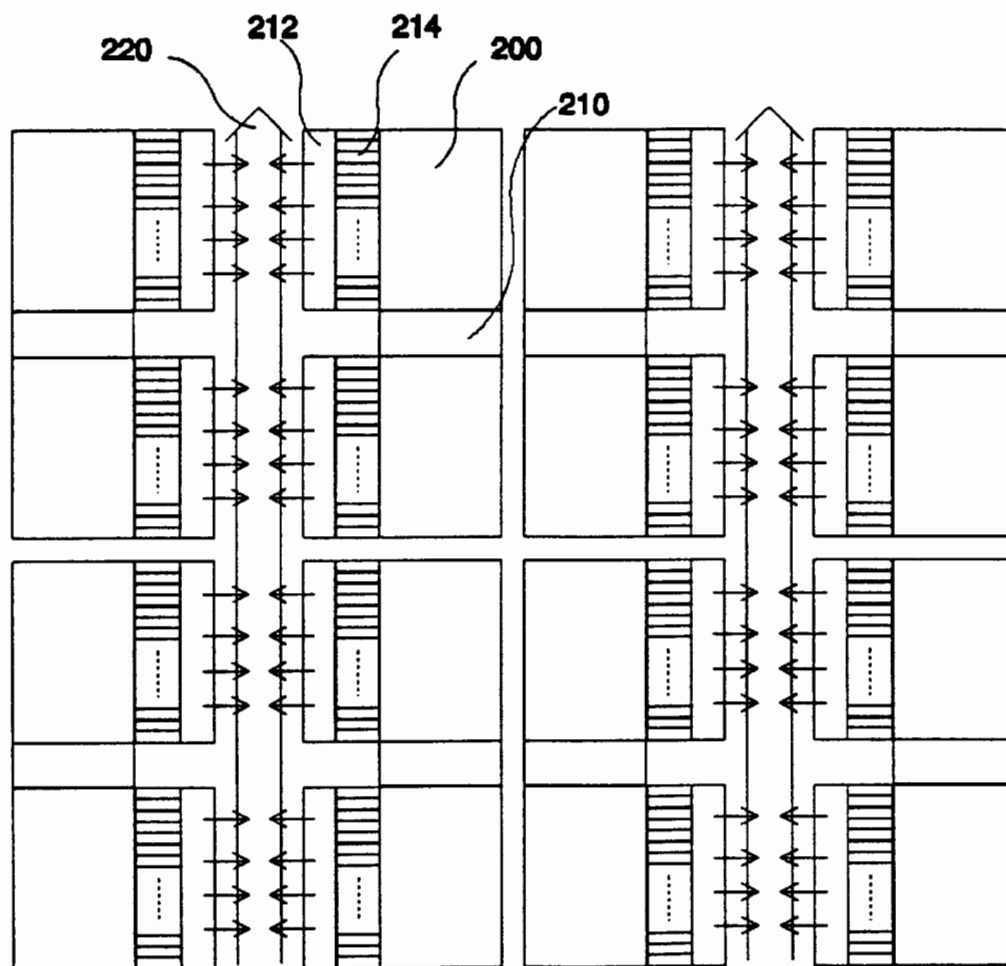


FIG. 3a

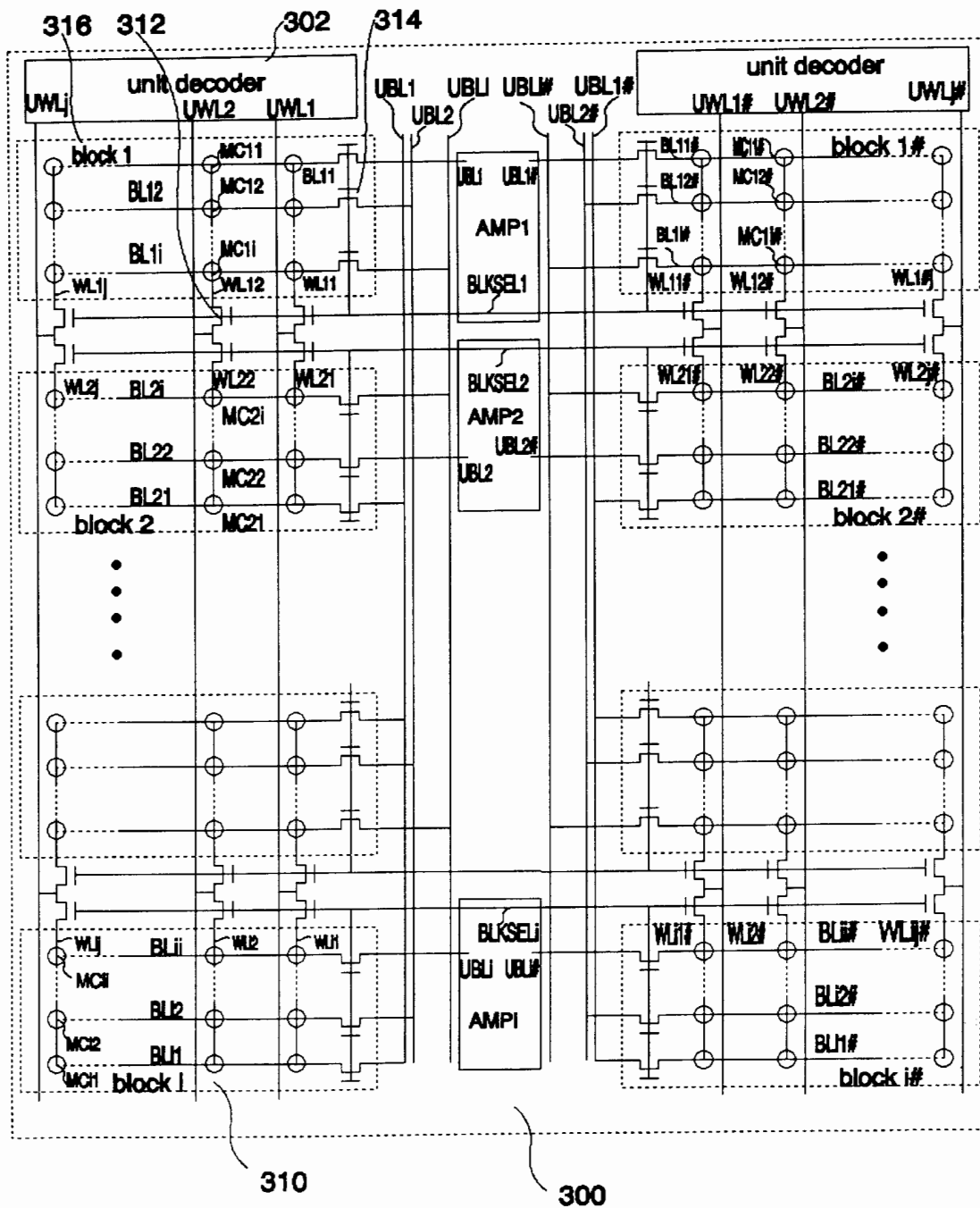
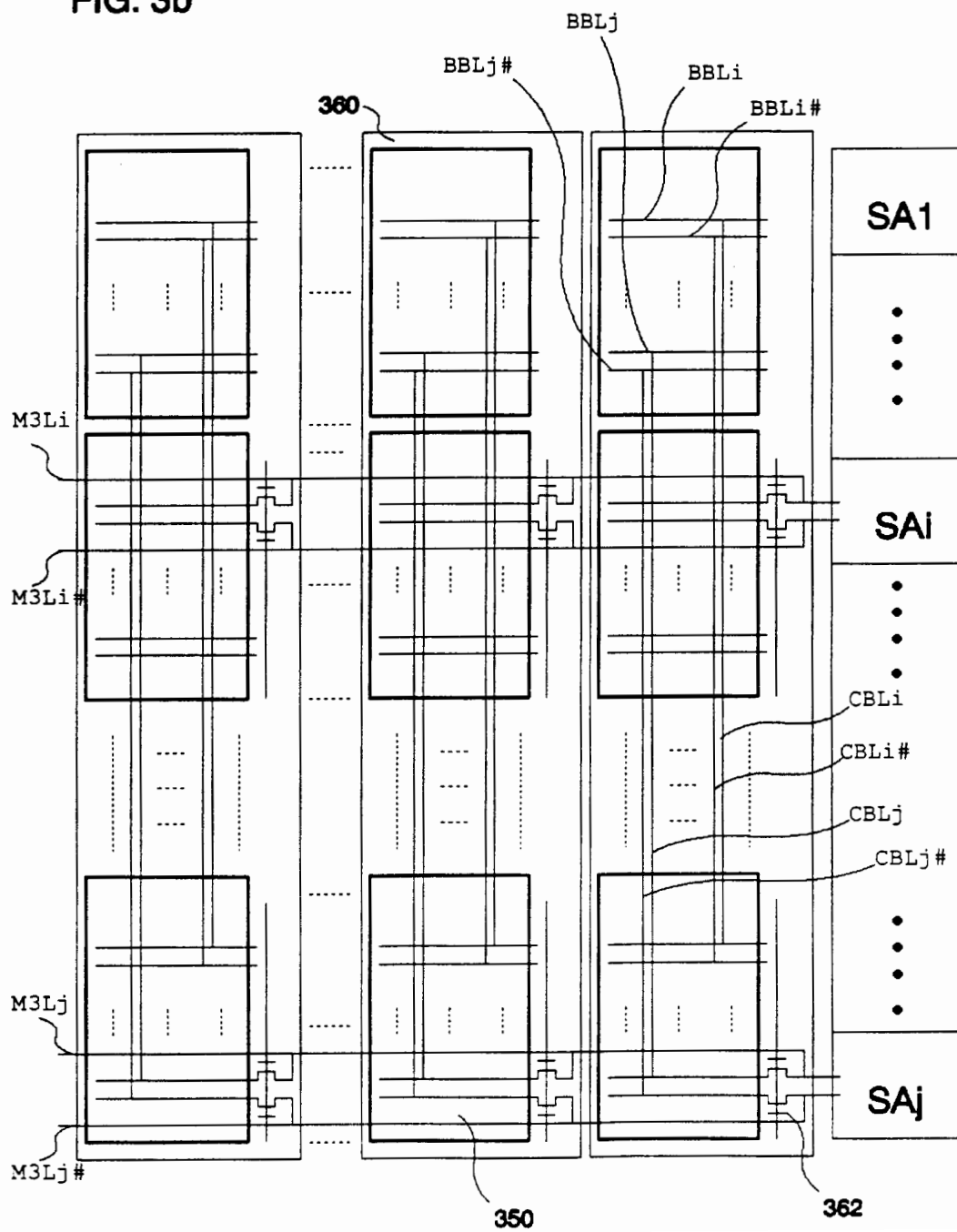




FIG. 3b



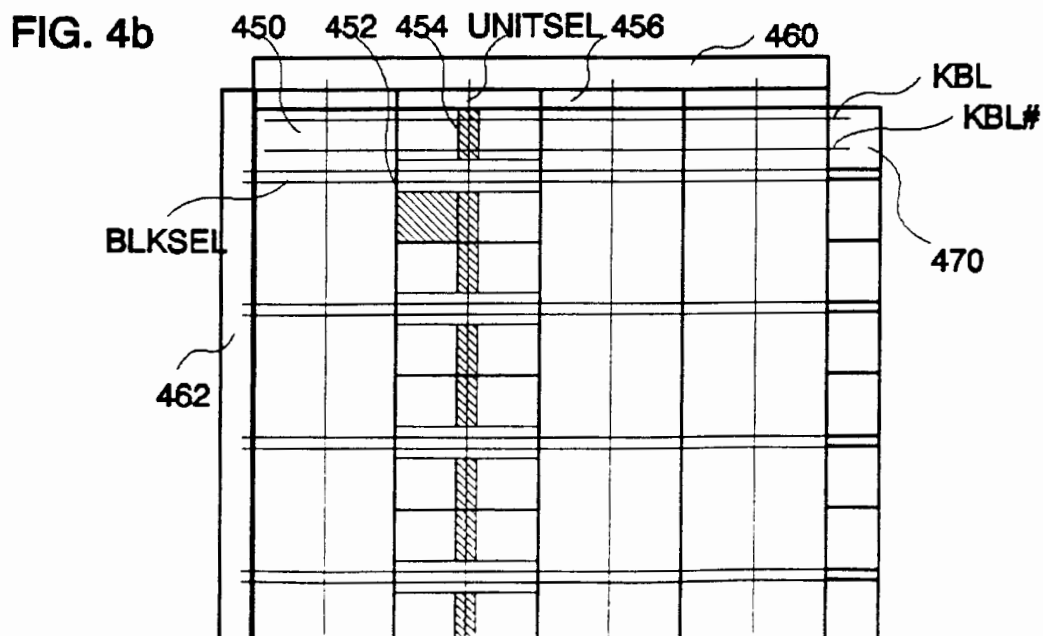
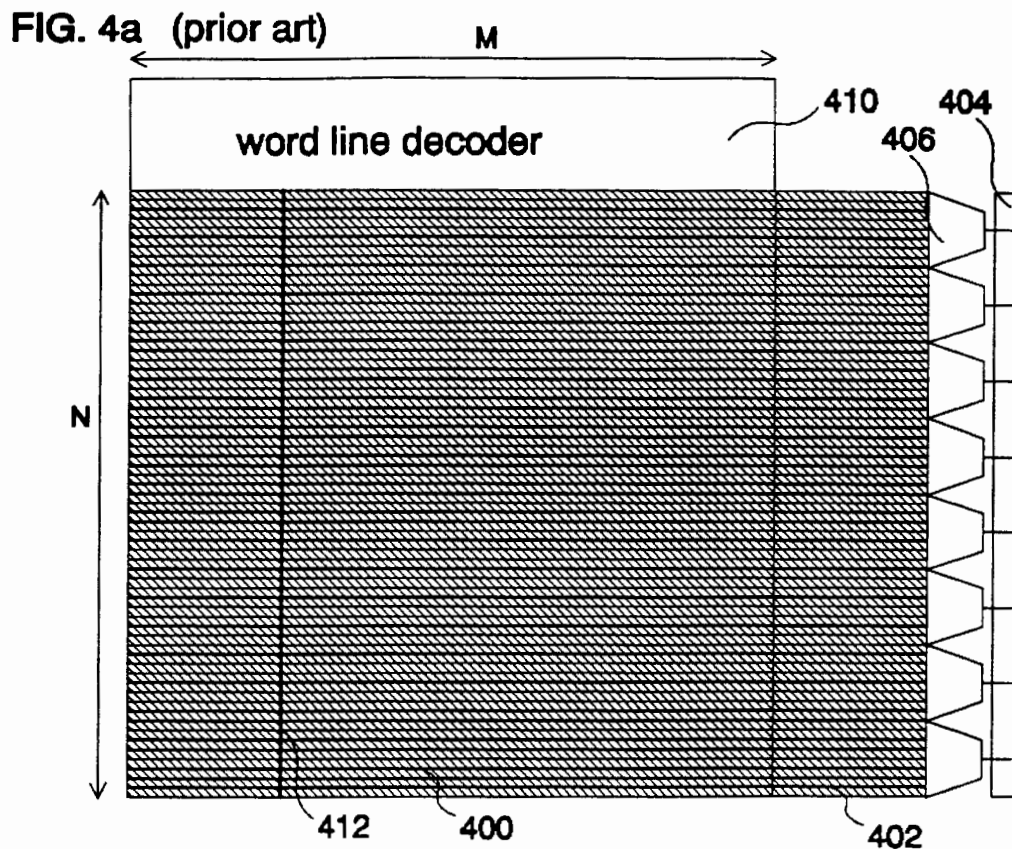


FIG. 5

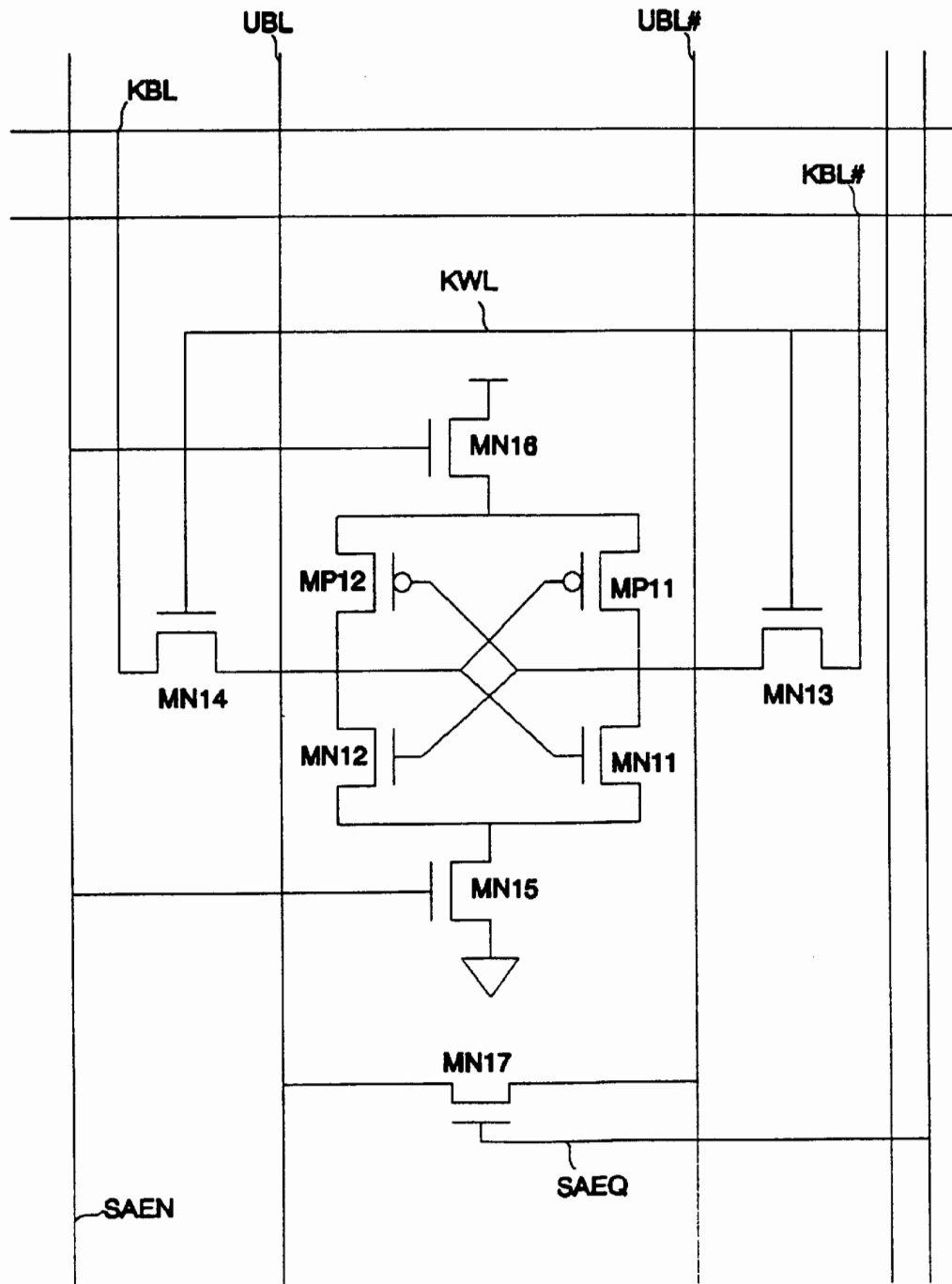


FIG. 6

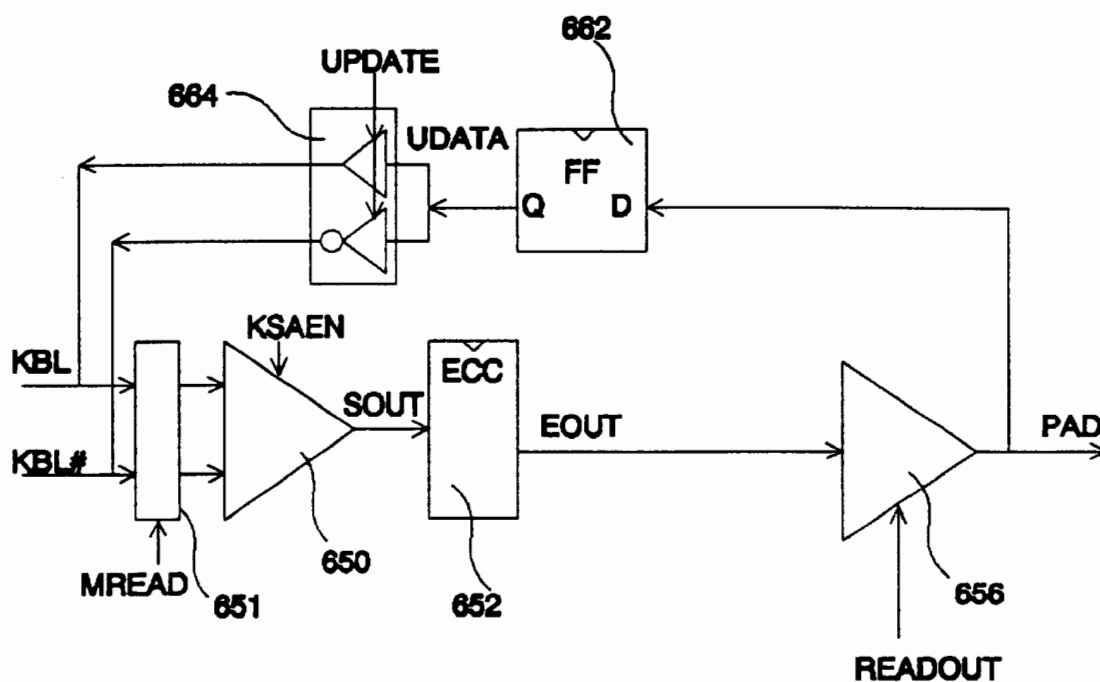


FIG. 7a (read cycle)

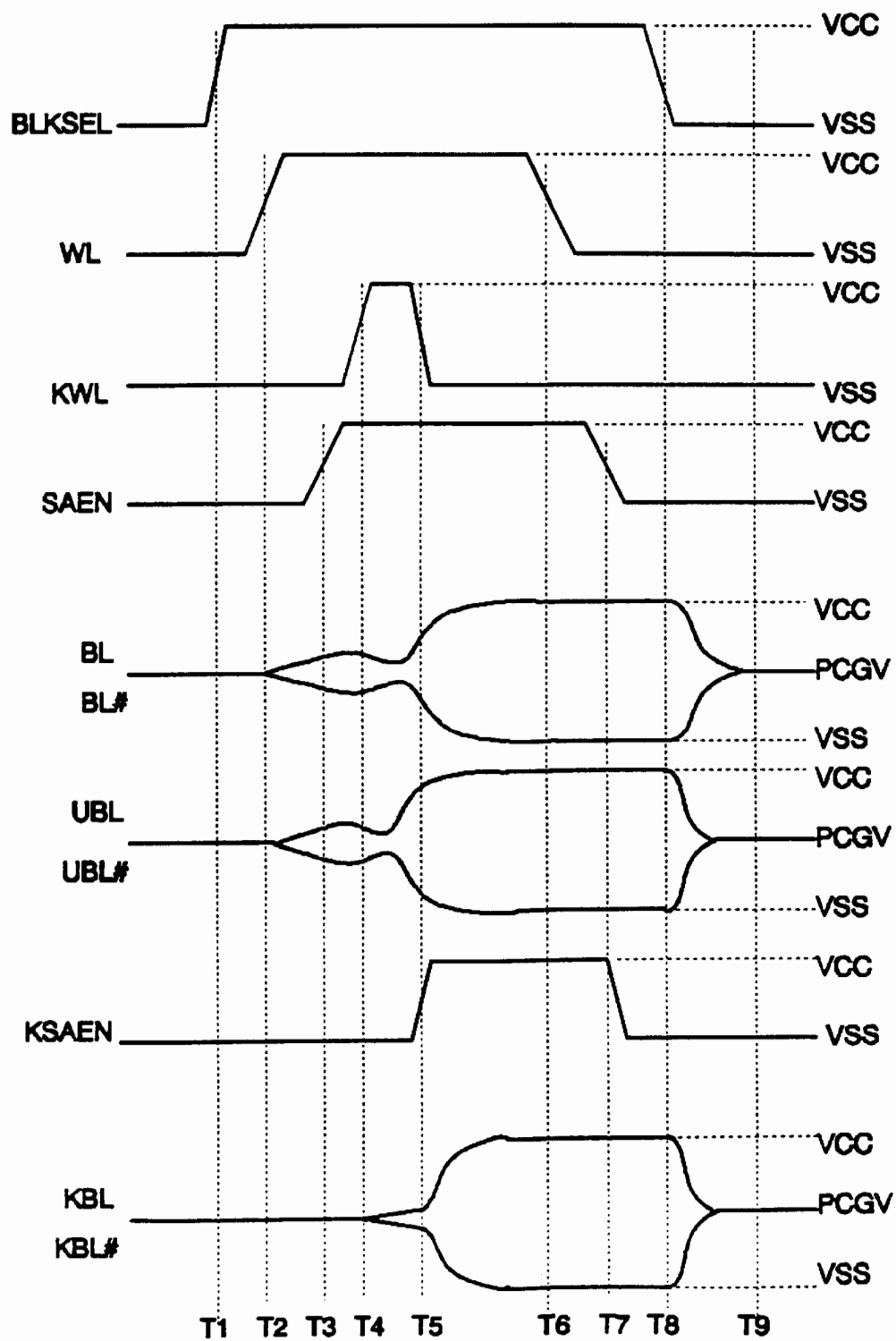




FIG. 7b (refresh cycle)

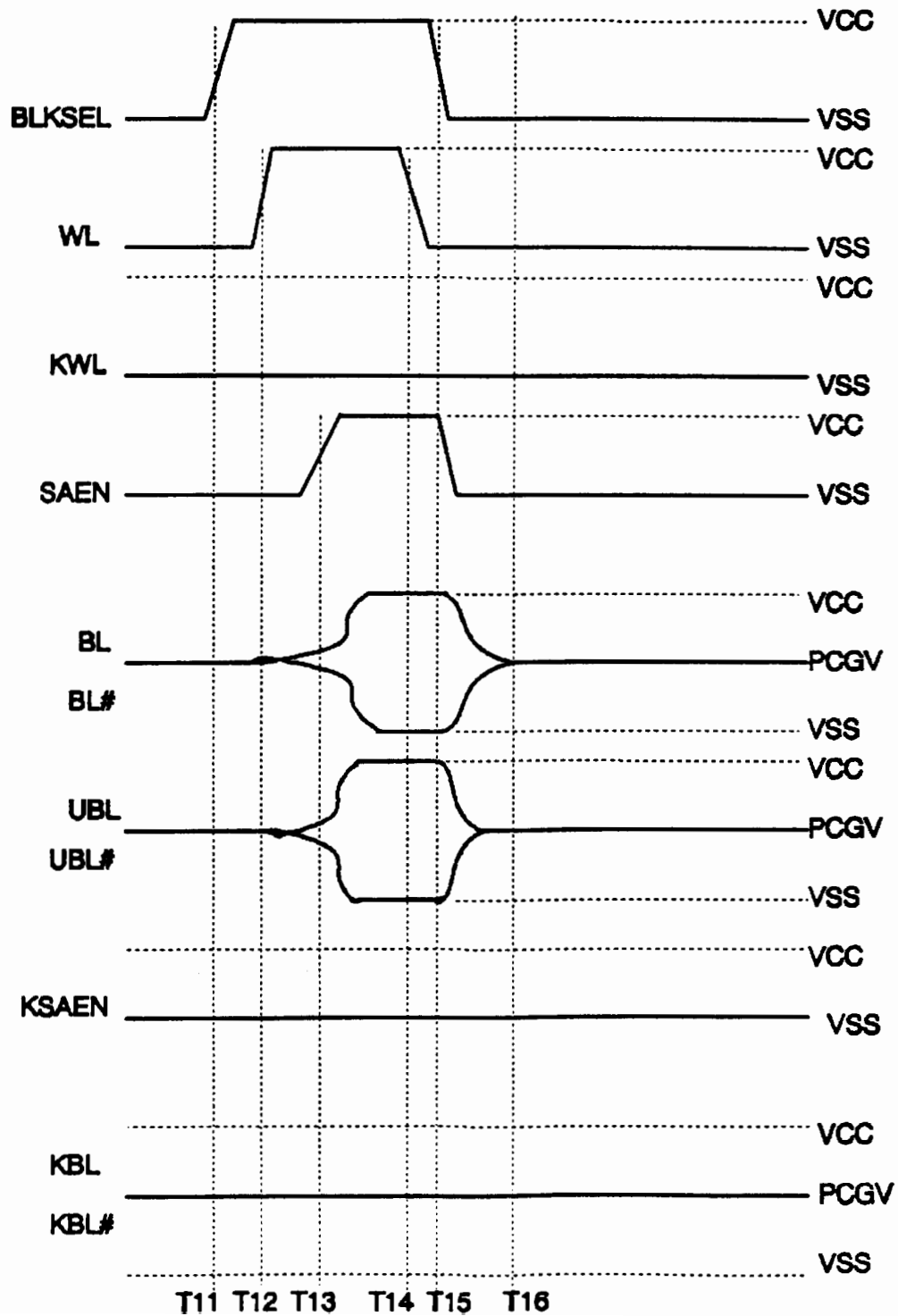


FIG. 7c (update cycle)

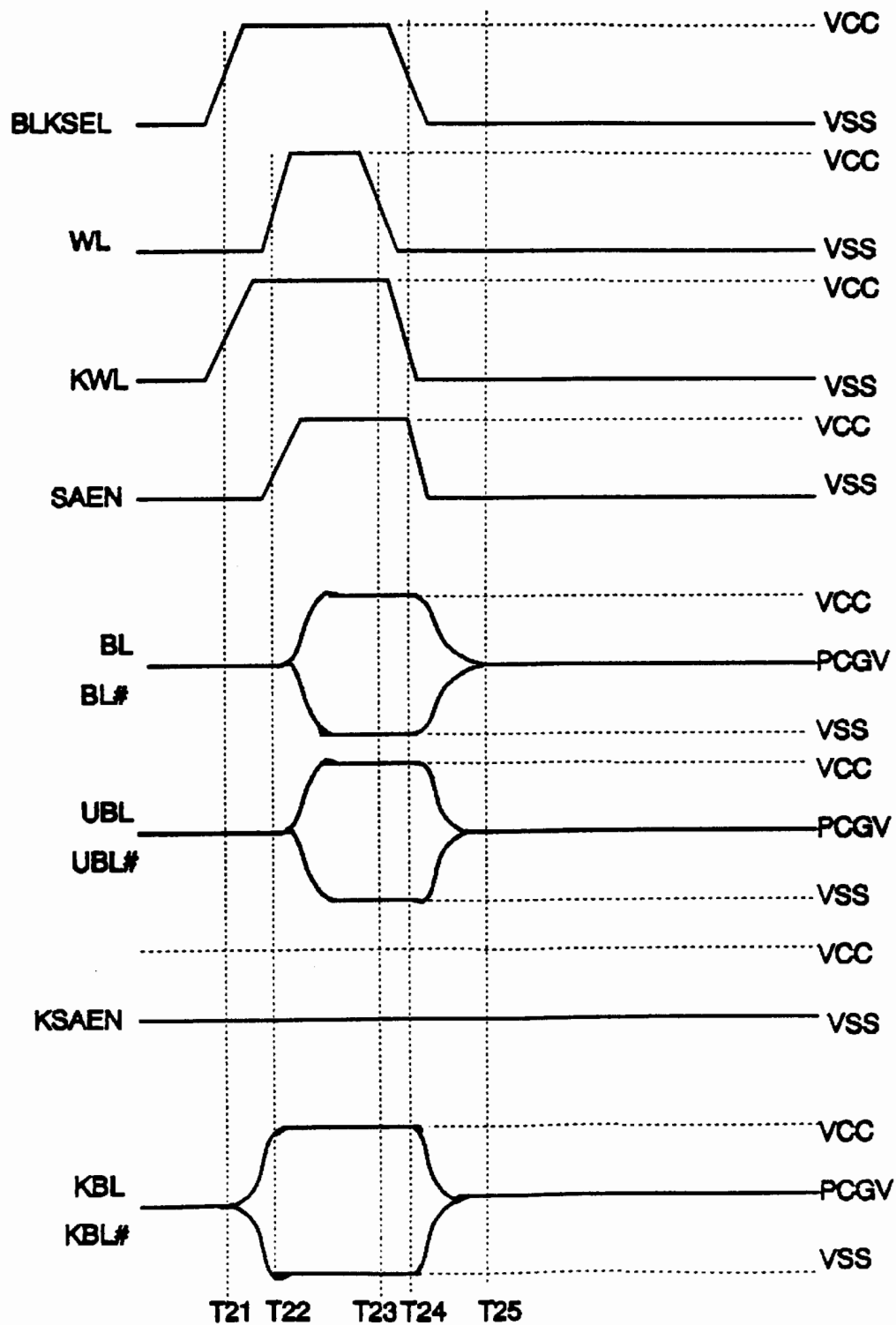


FIG. 8

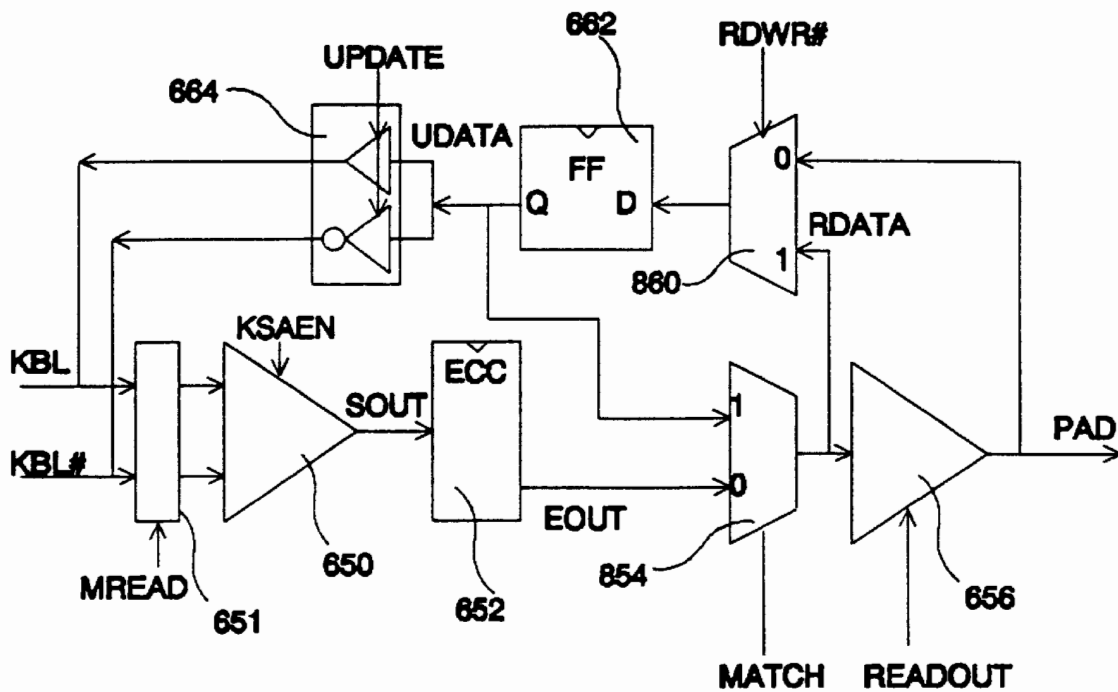
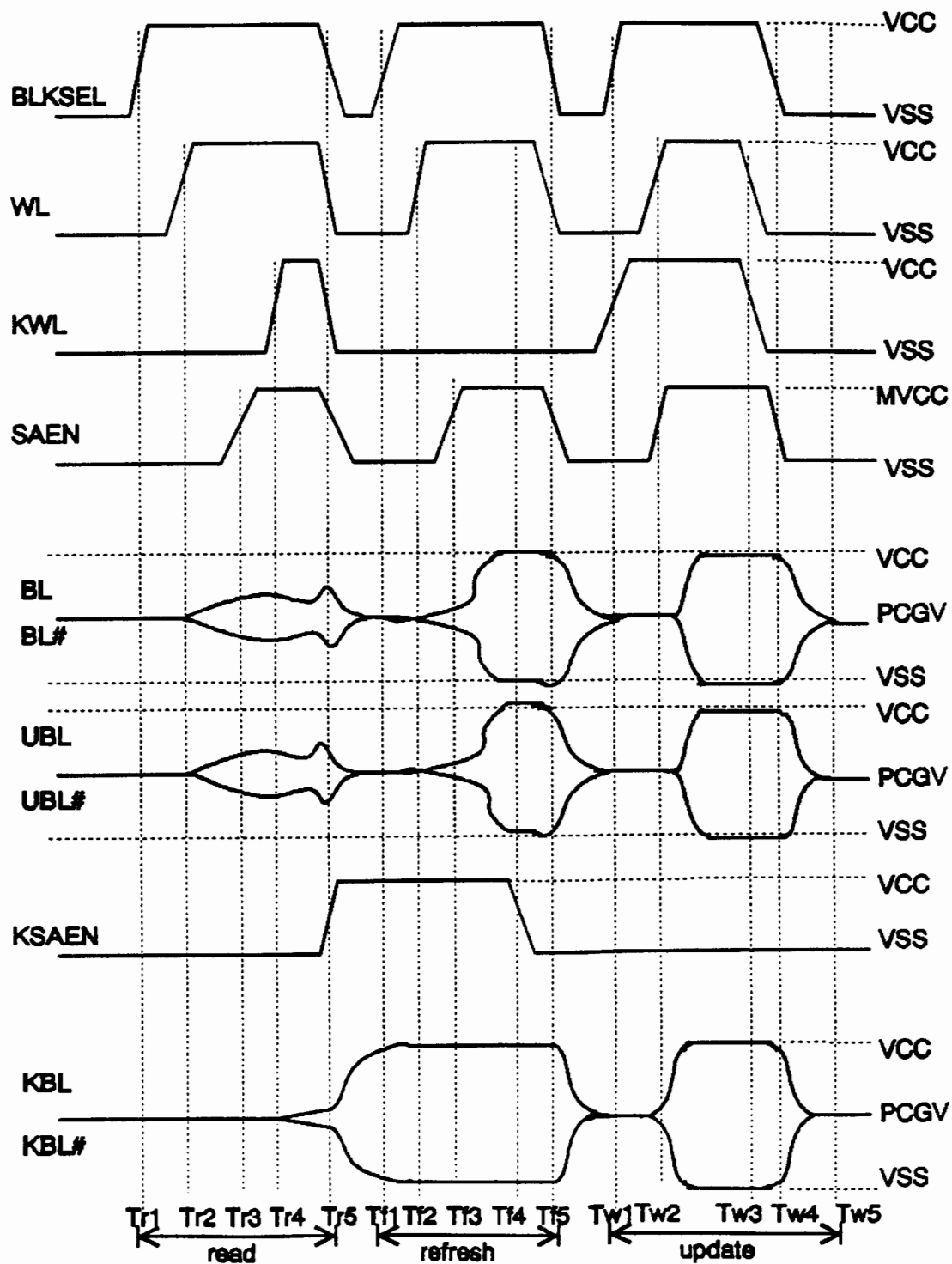


FIG. 9



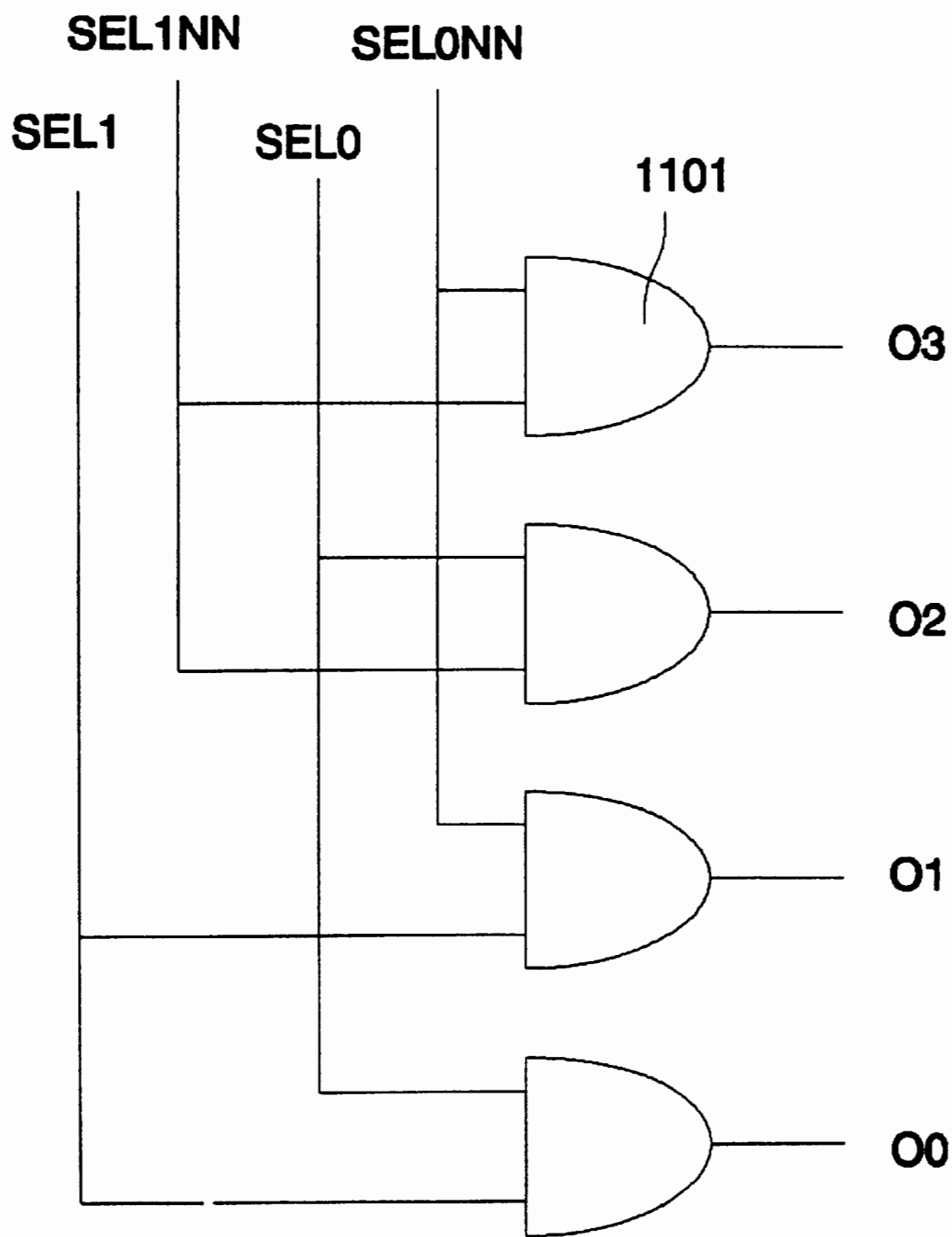
**FIG. 10**



FIG. 11(a)

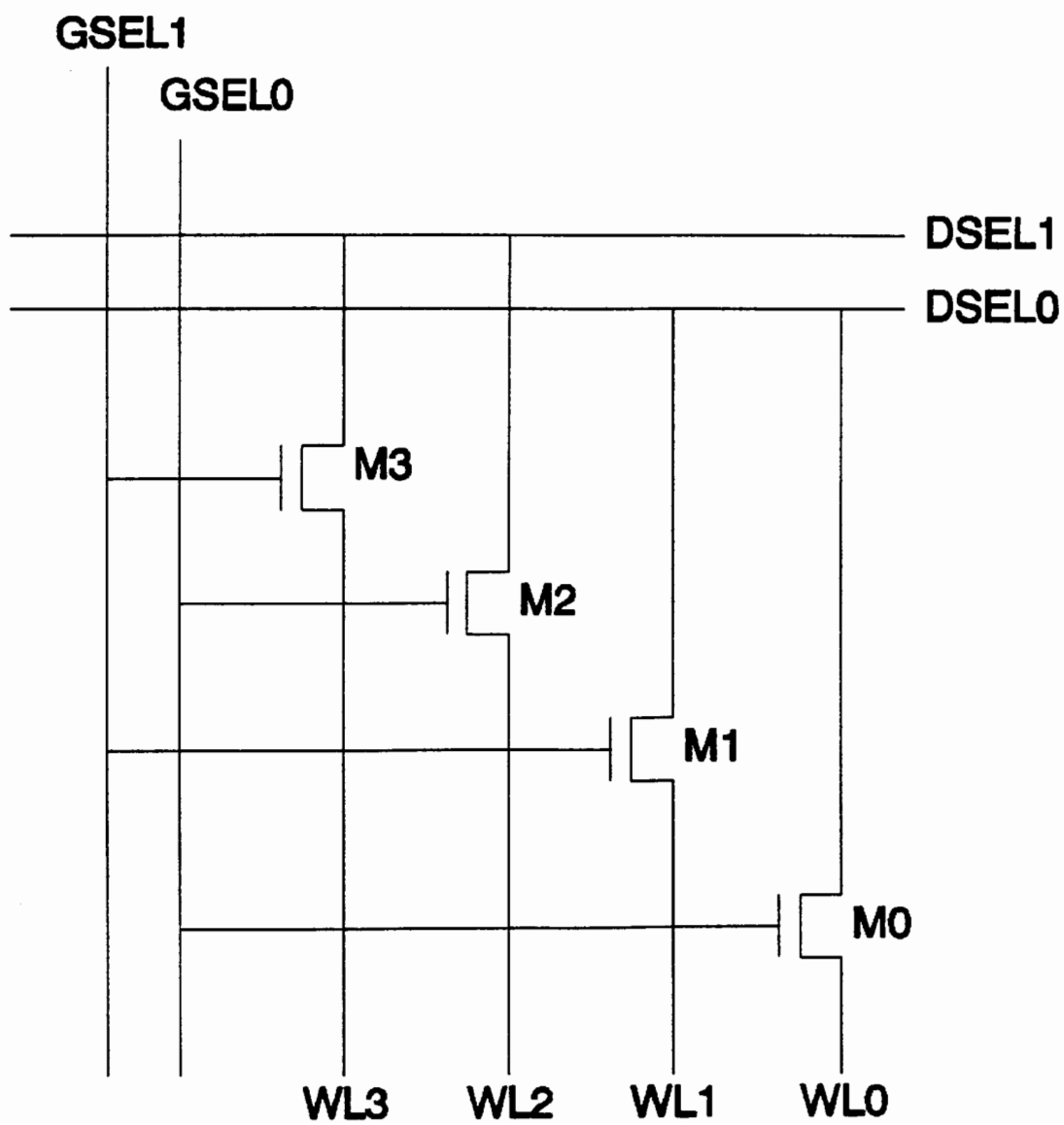


FIG. 11(b)

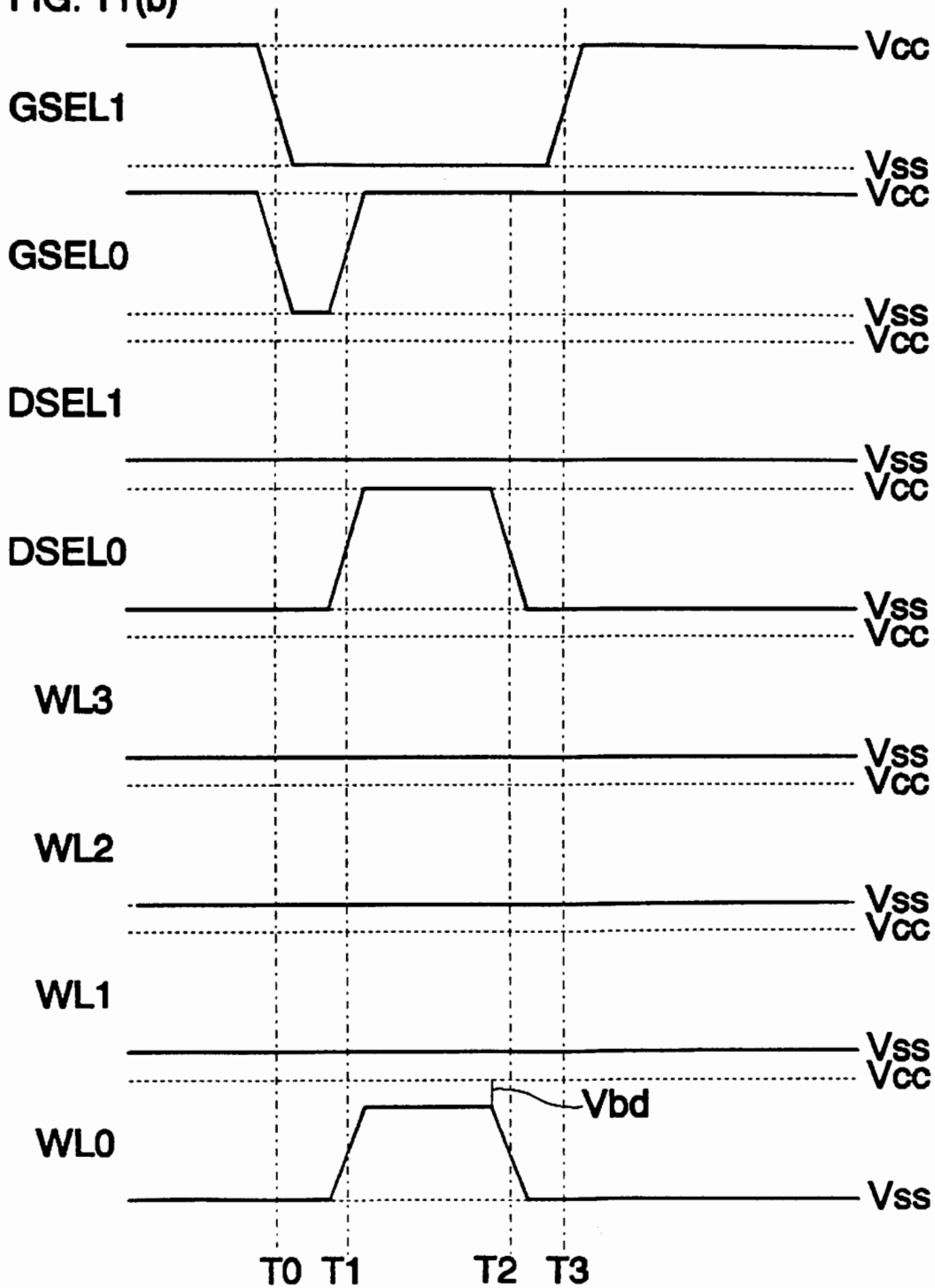


FIG. 12(a)

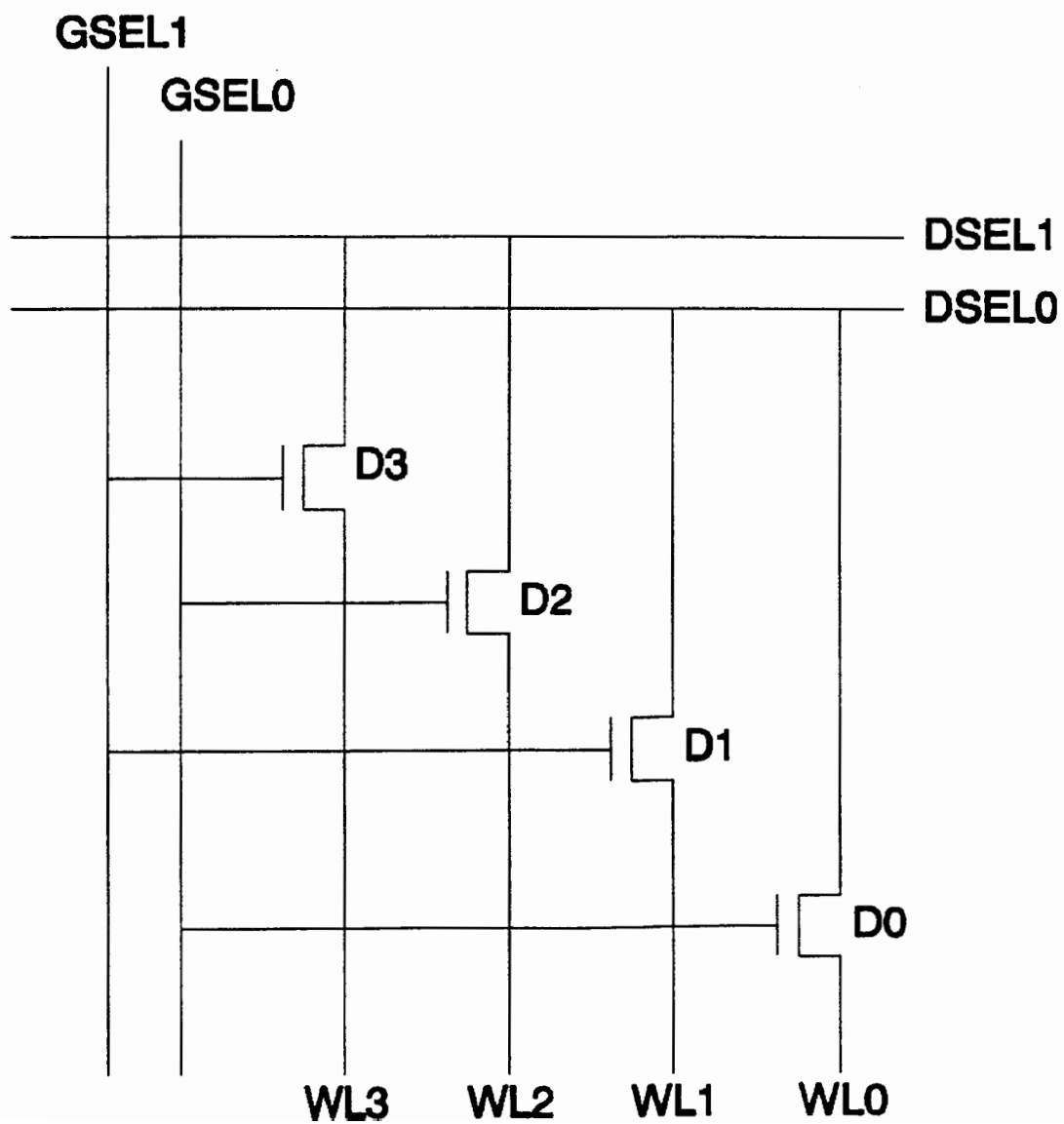


FIG. 12(b)

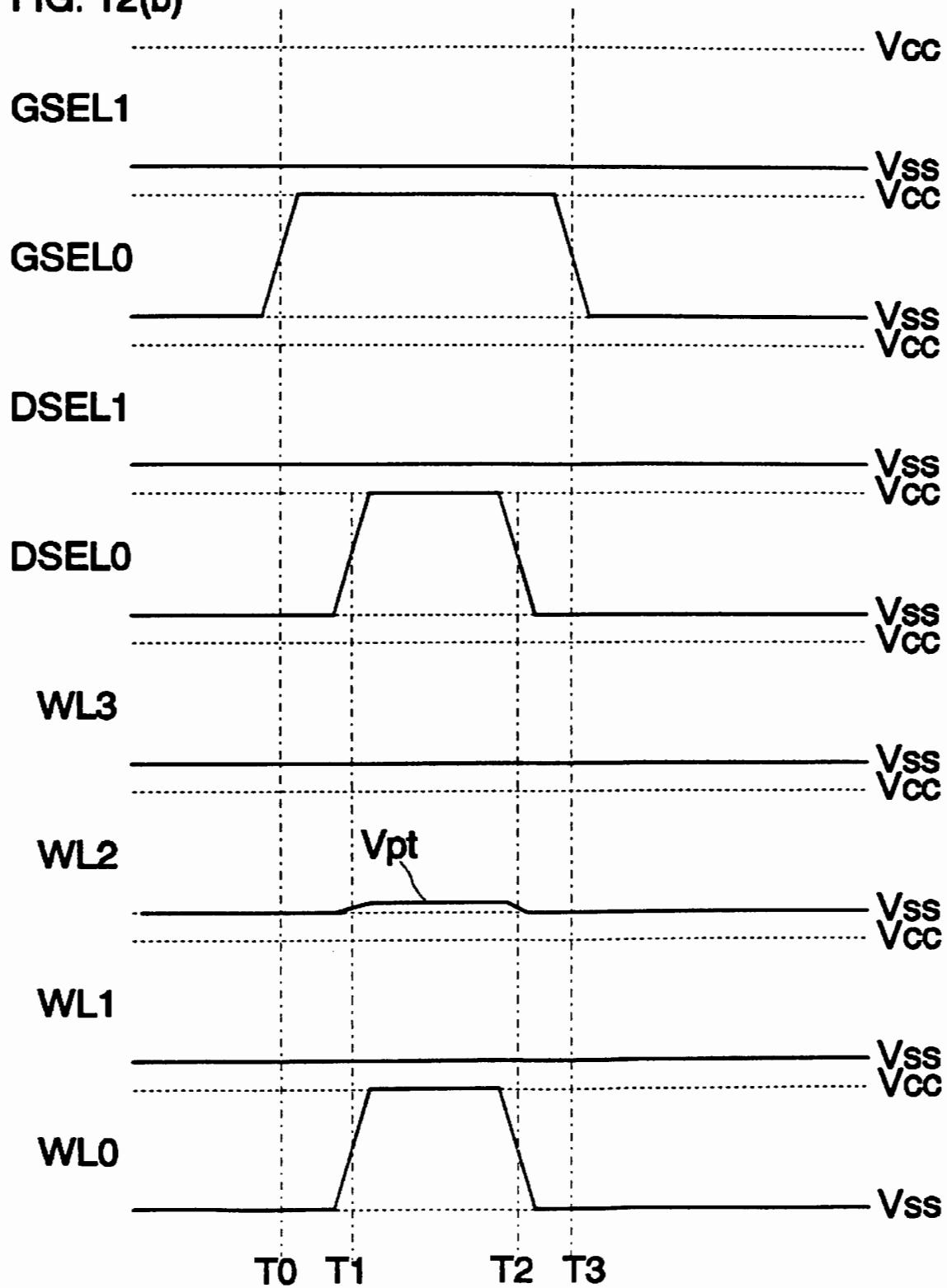


FIG. 12(c)

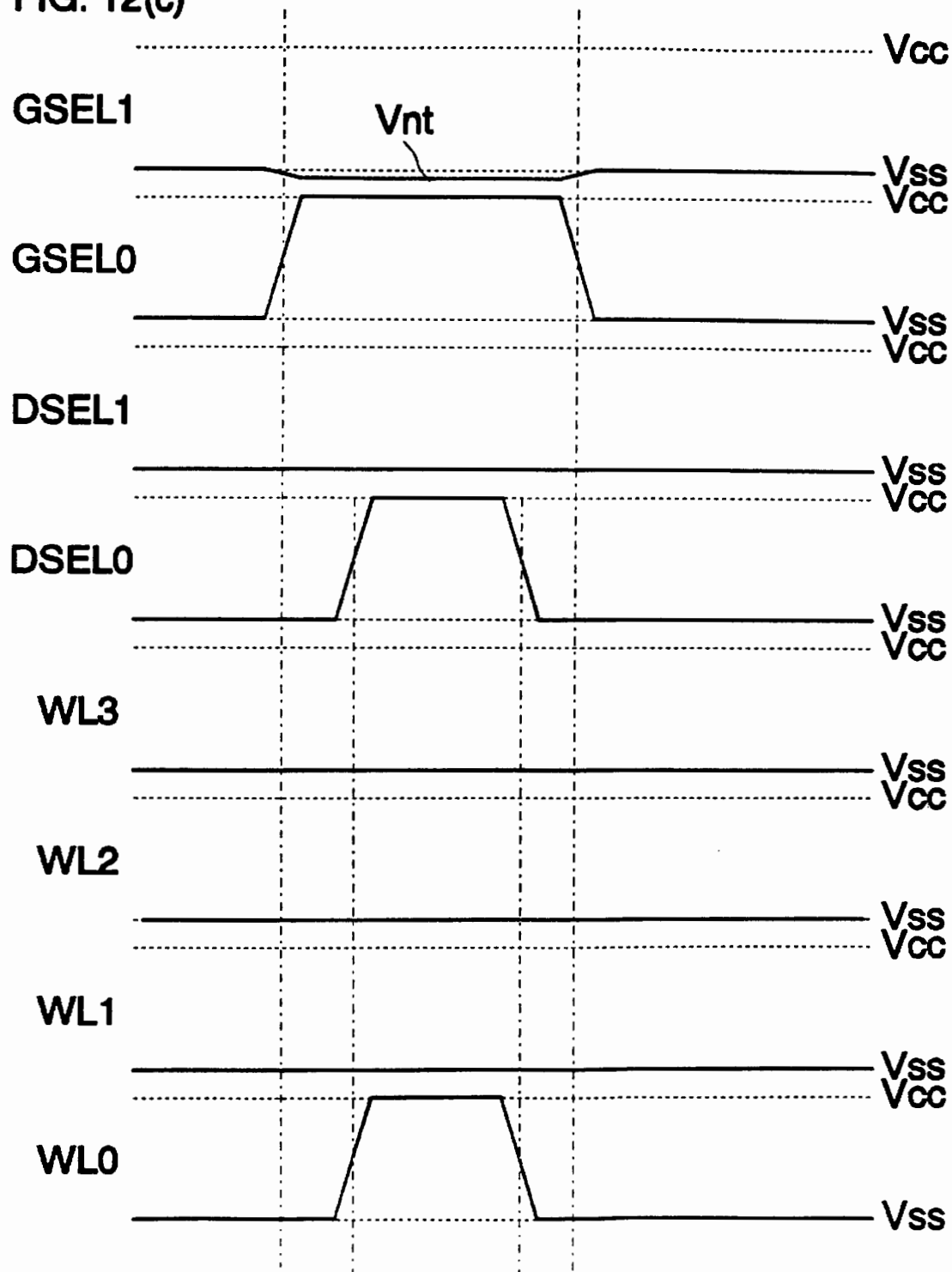




FIG. 13

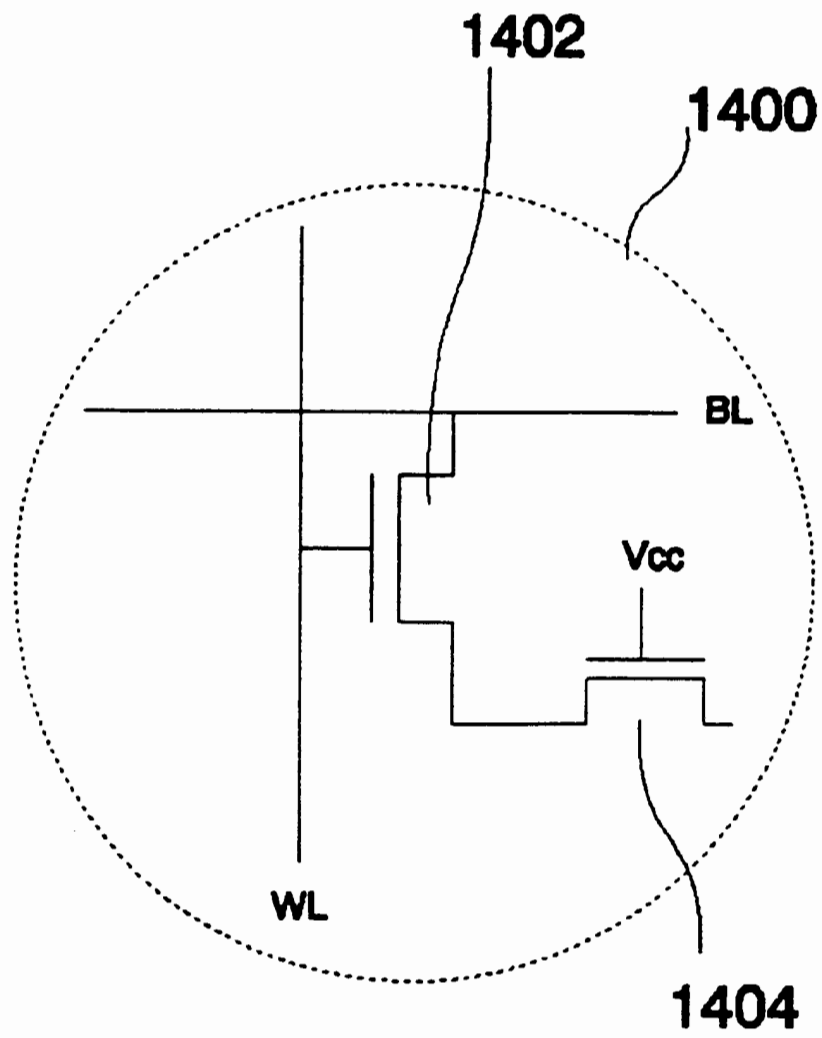


FIG. 14(a)

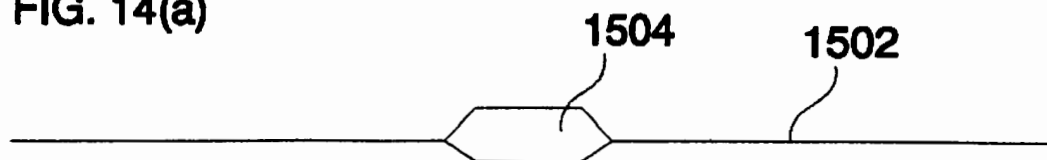


FIG. 14(b)

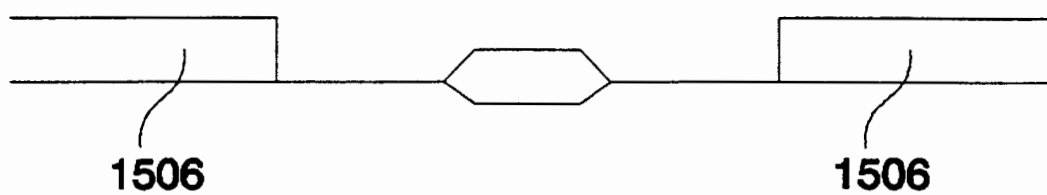


FIG. 14(c)

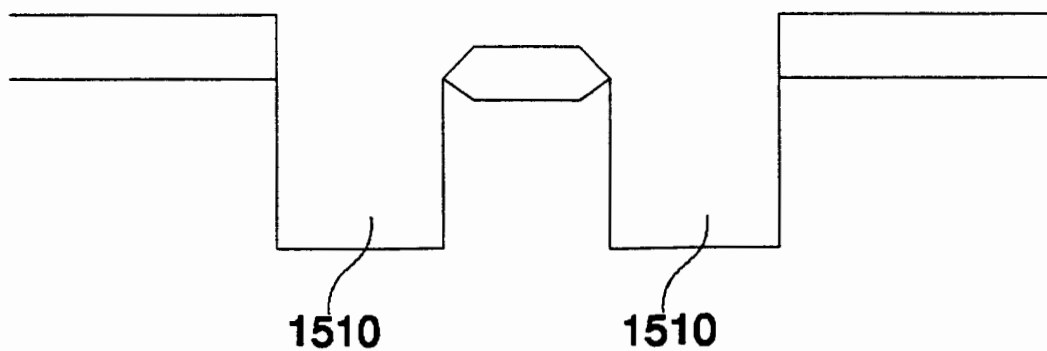


FIG. 14(d)

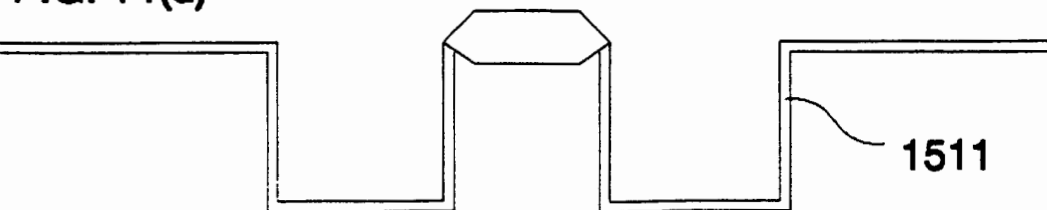


FIG. 14(e)

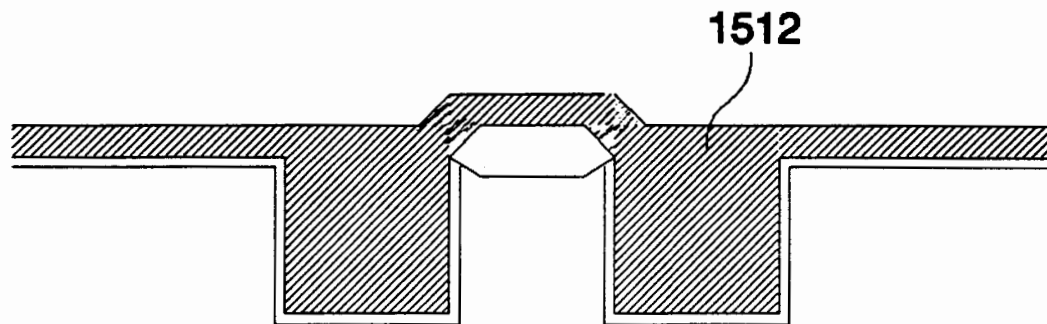


FIG. 14(f)

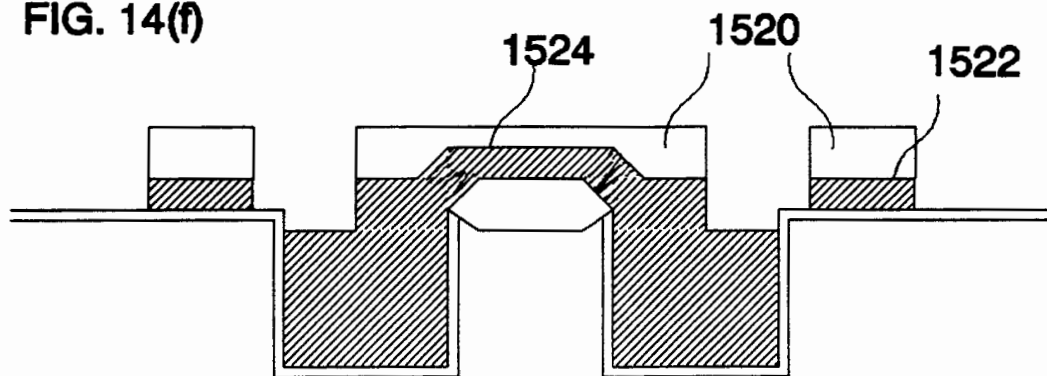
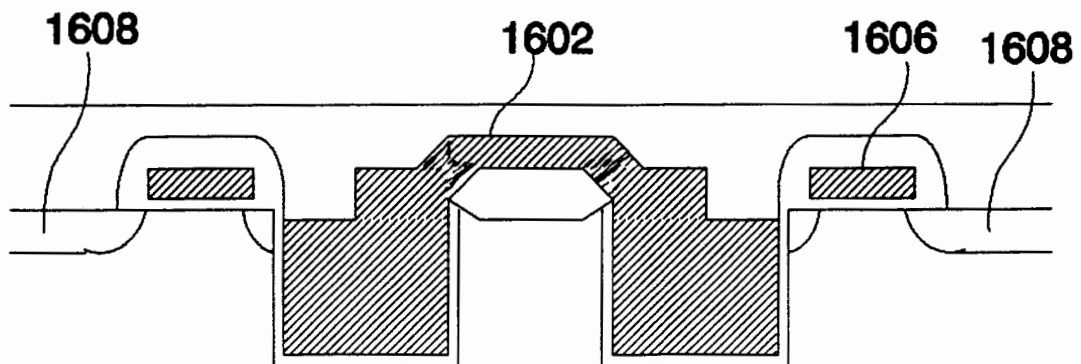


FIG. 14(g)



**FIG. 15(a)**

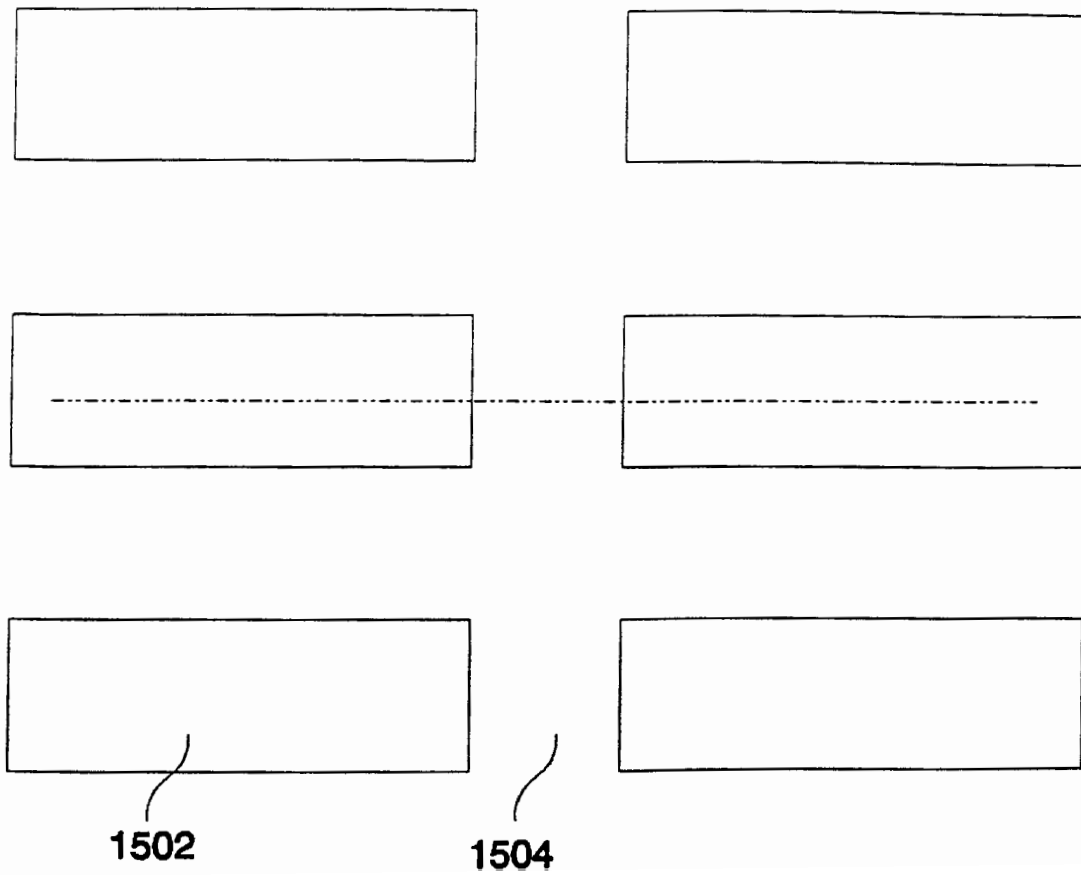


FIG. 15(b)

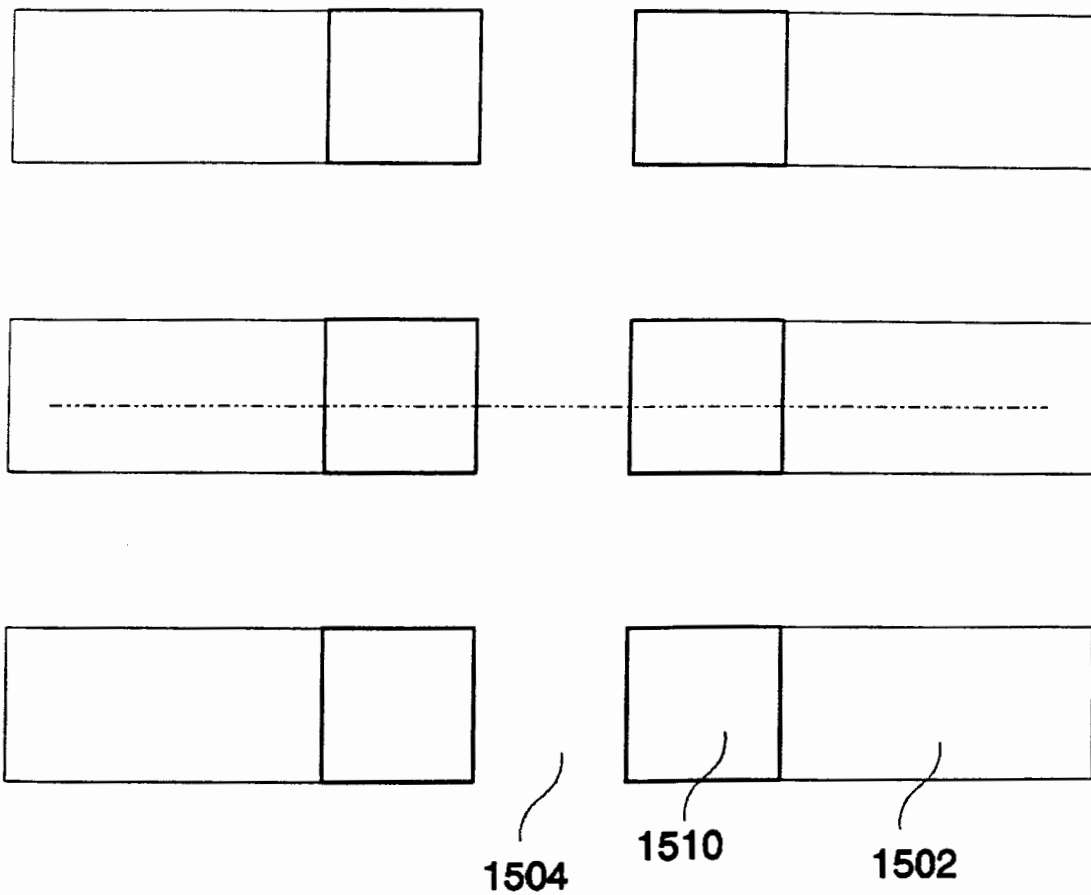




FIG. 15(c)

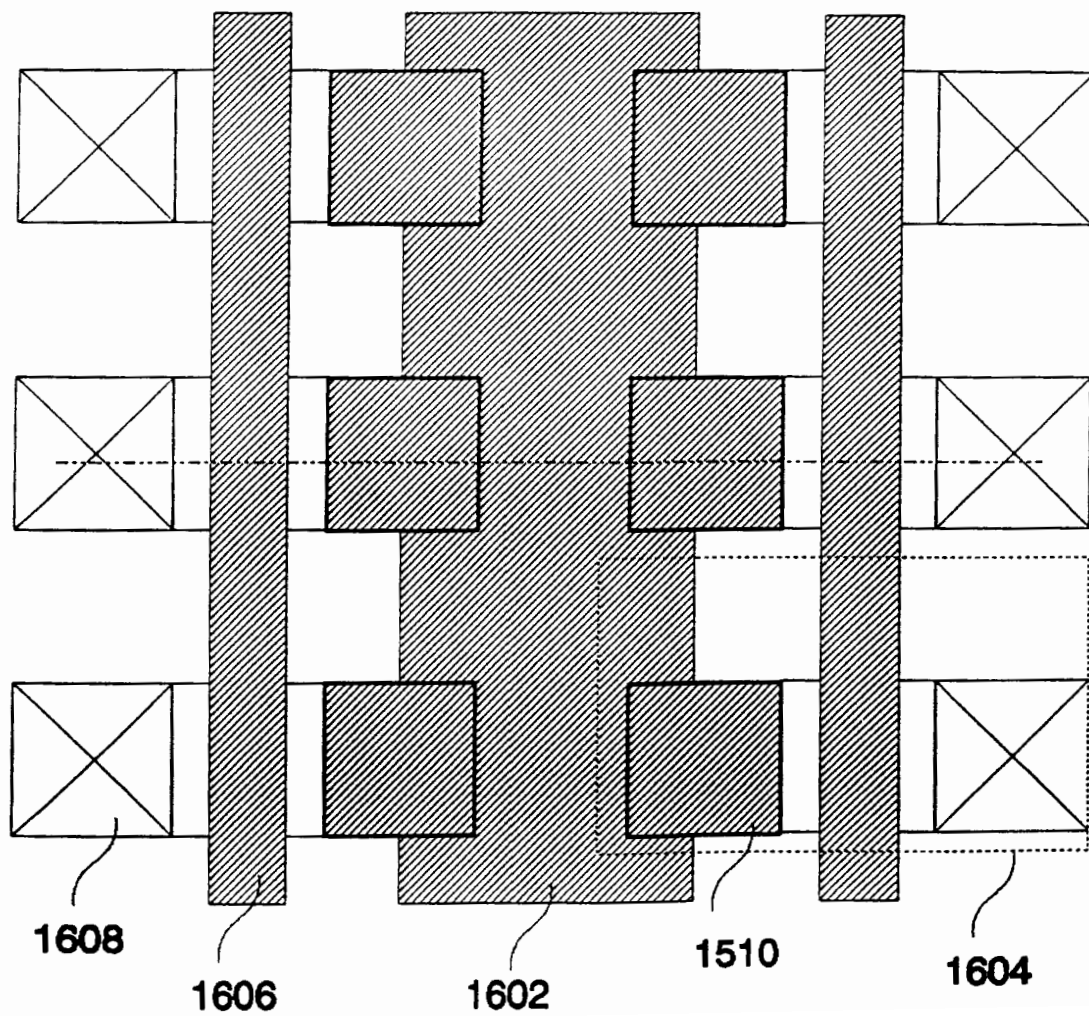


FIG. 16(a)

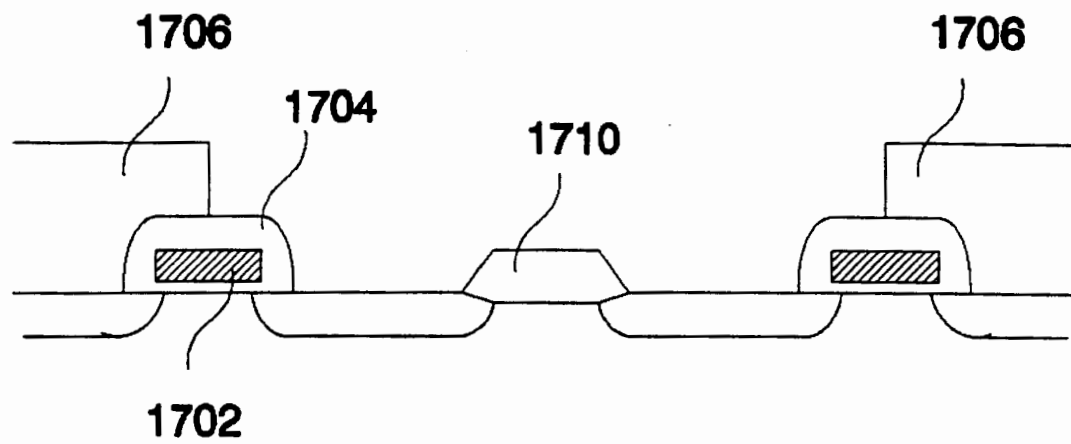


FIG. 16(b)

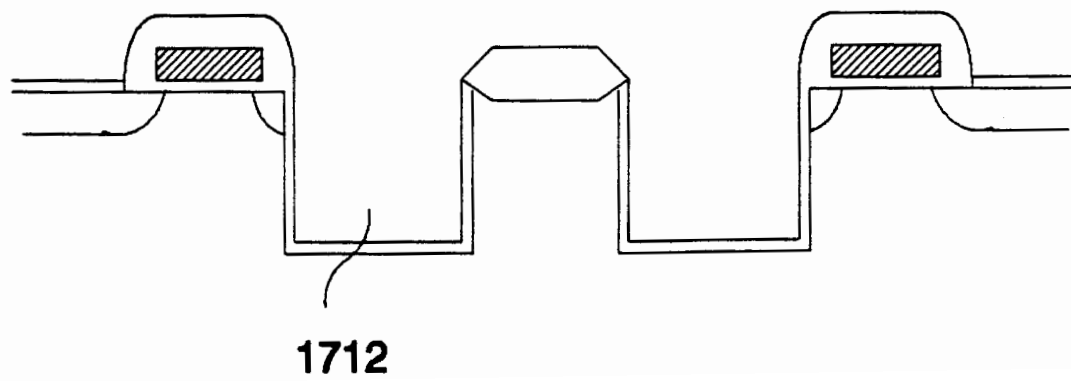


FIG. 16(c)

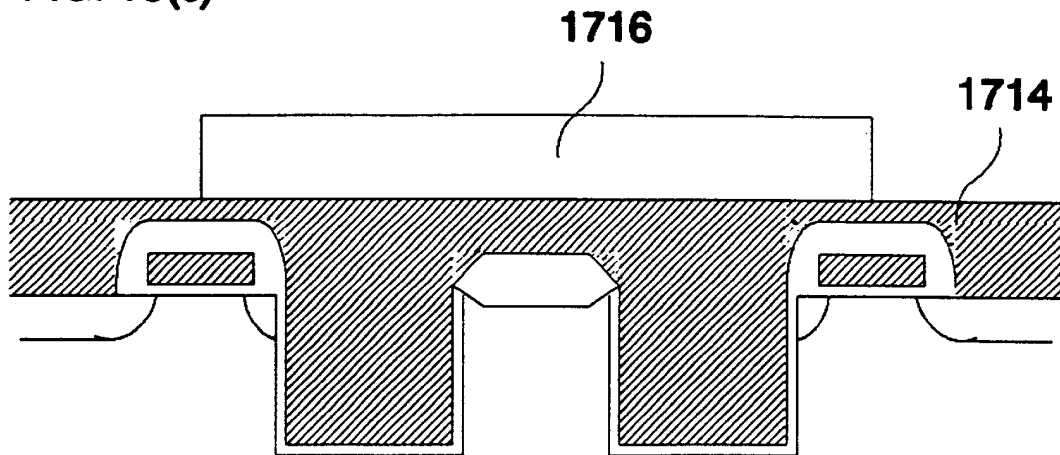


FIG. 16(d)

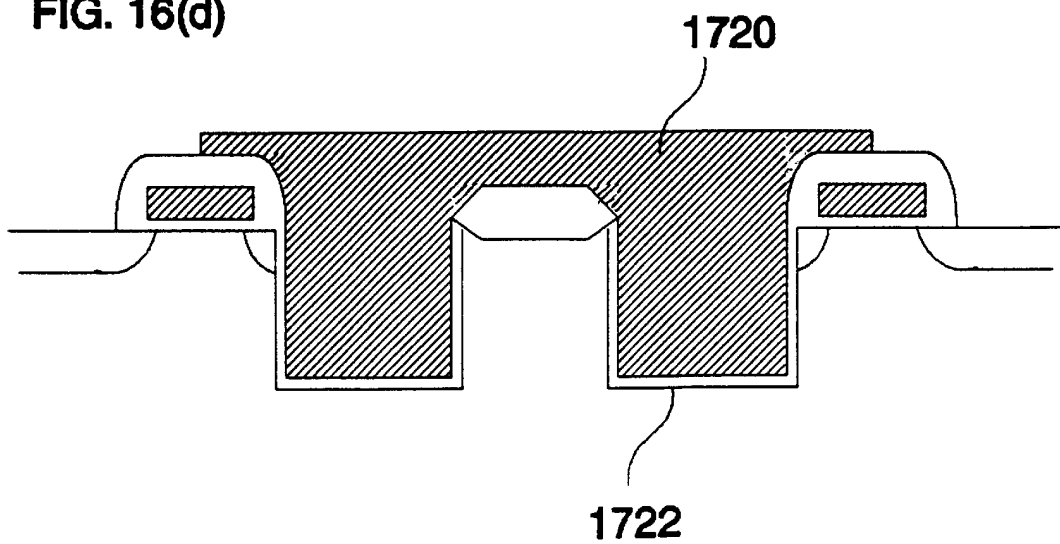


FIG. 17

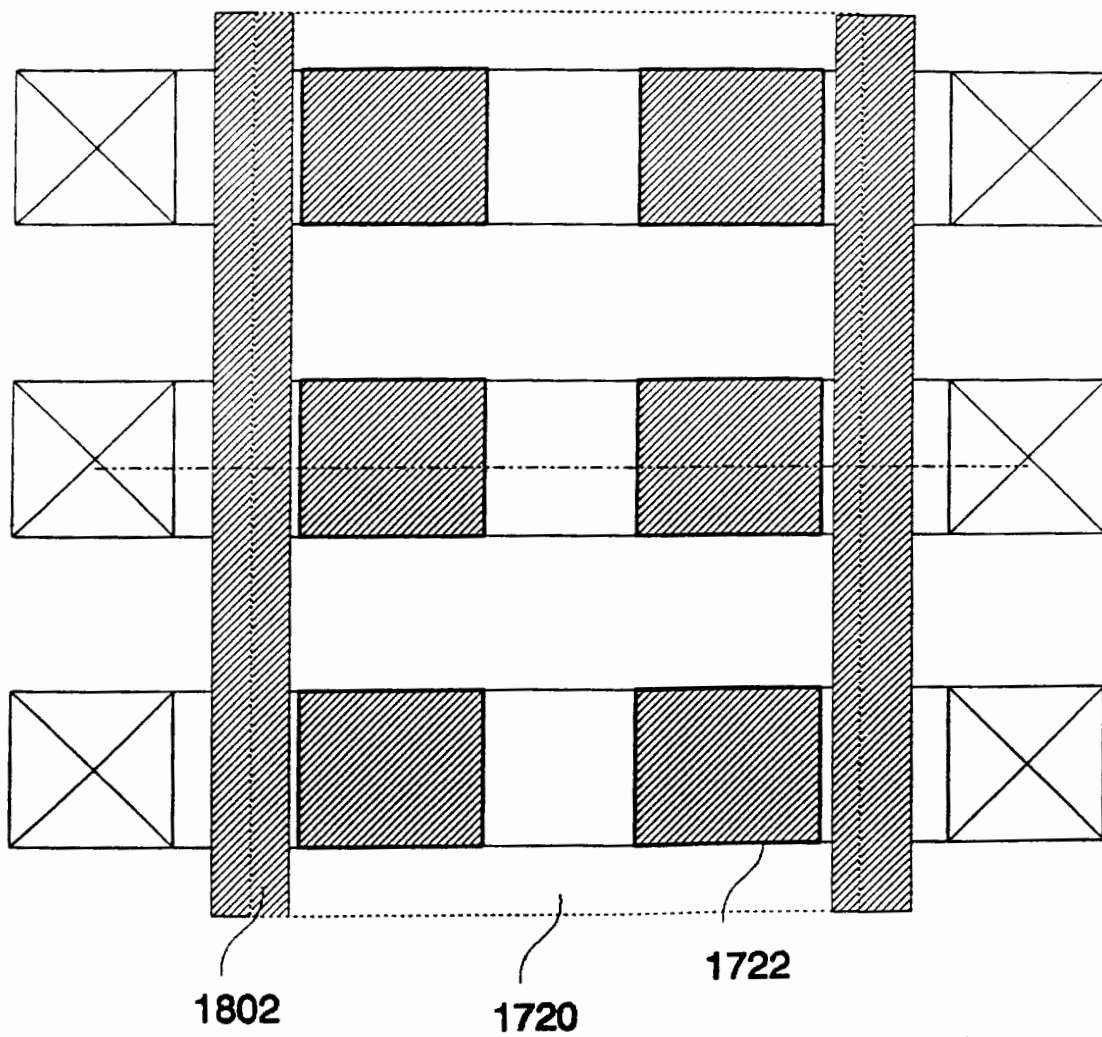


FIG. 18(a)

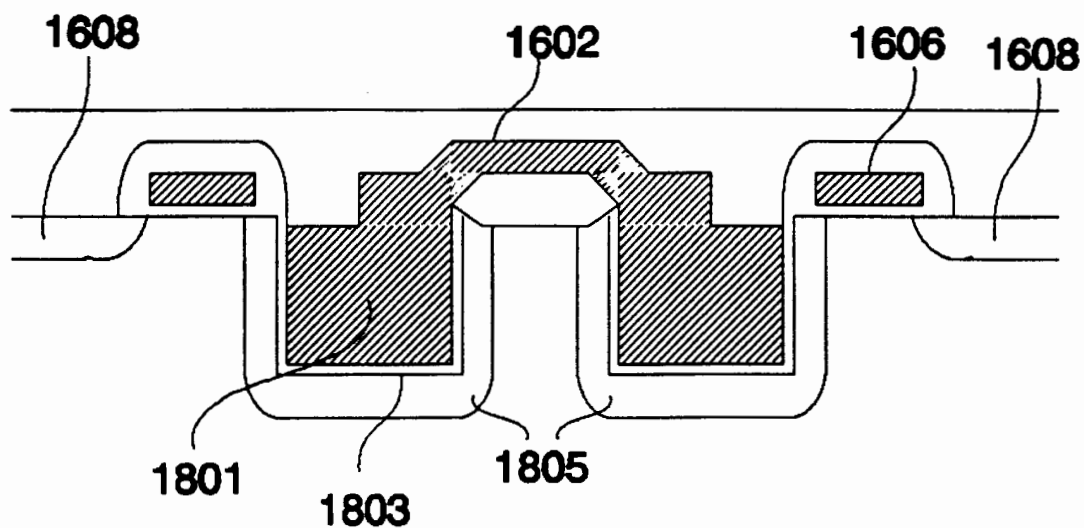


FIG. 18(b)

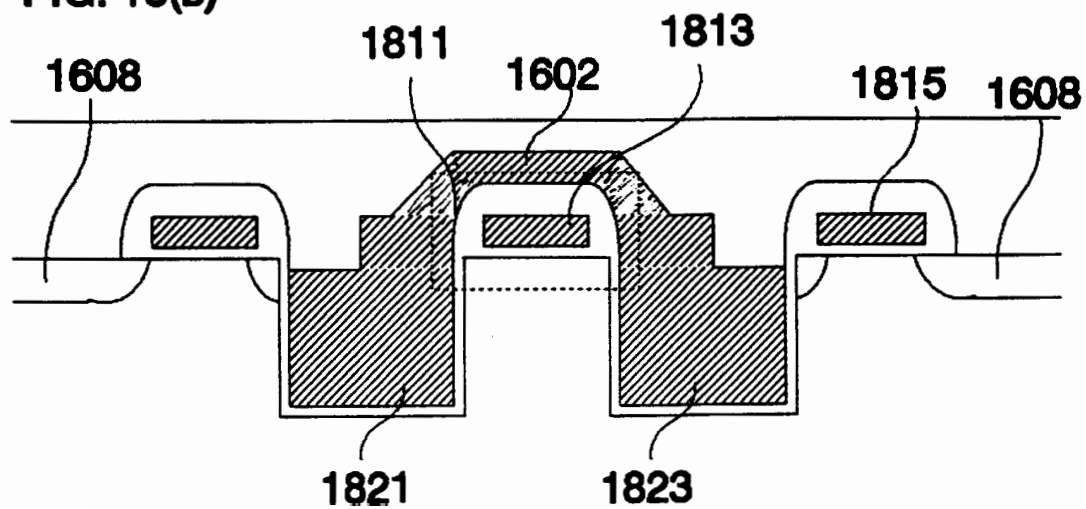
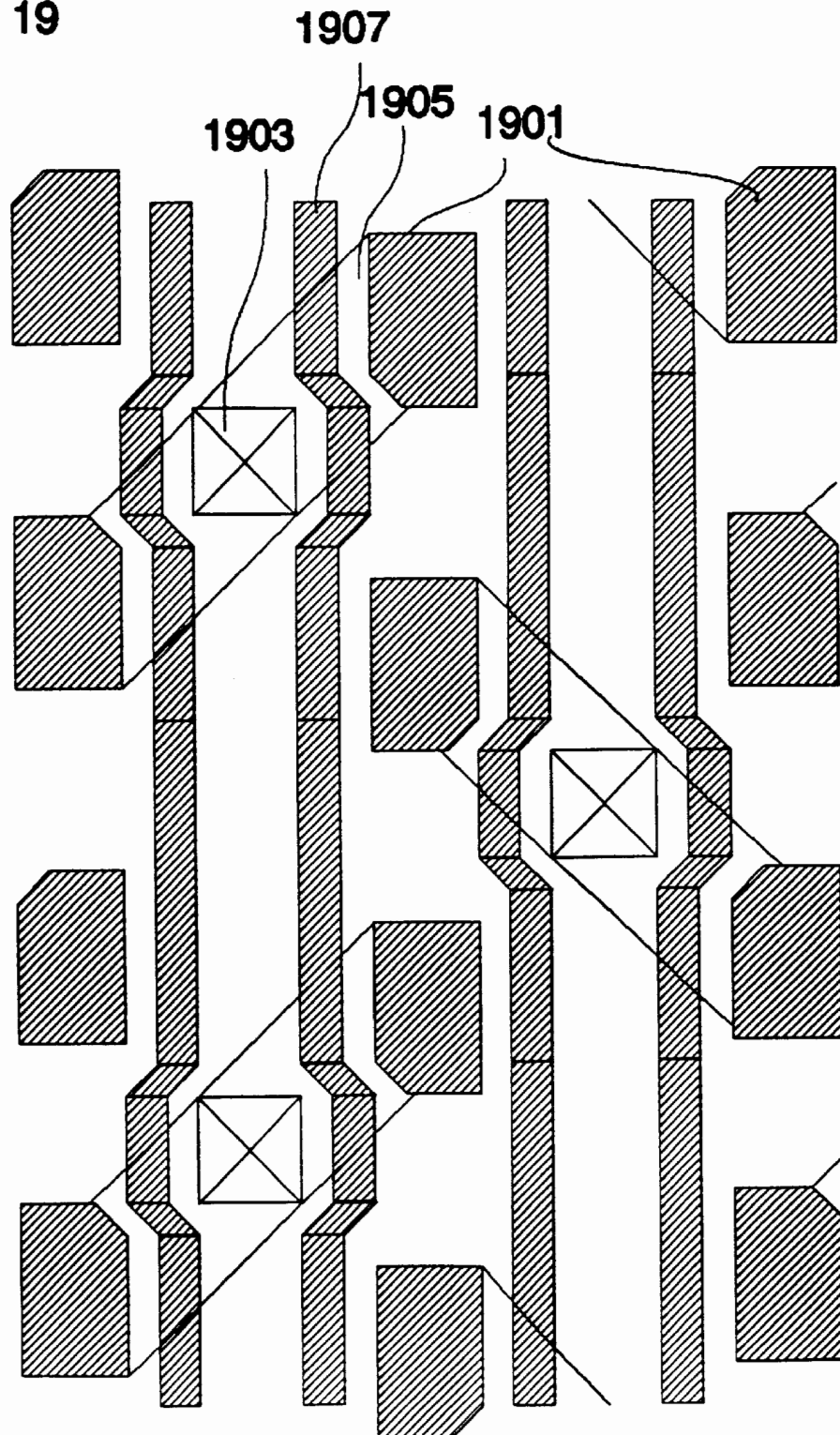
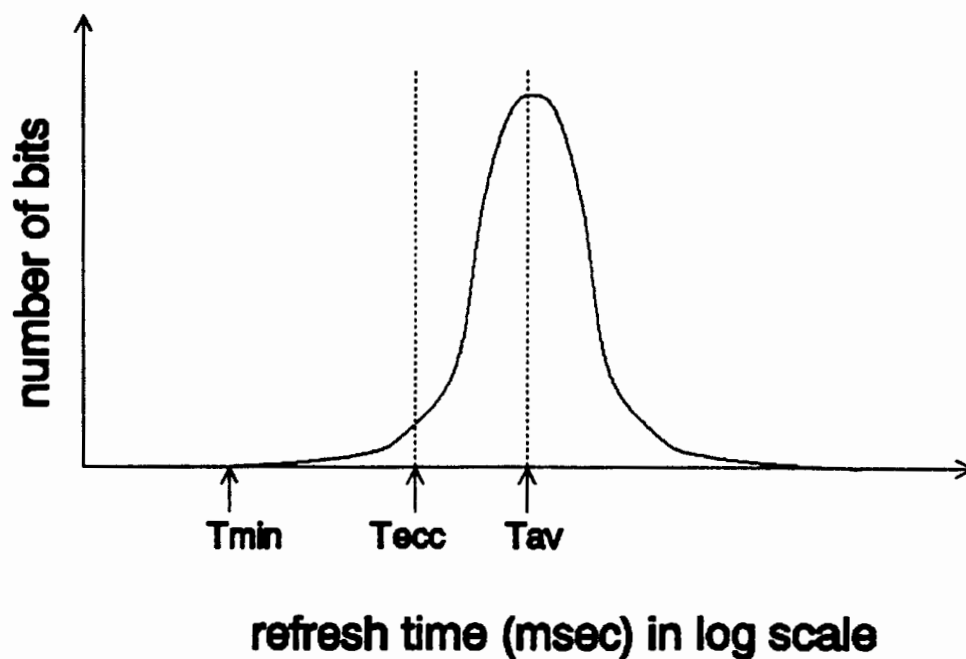
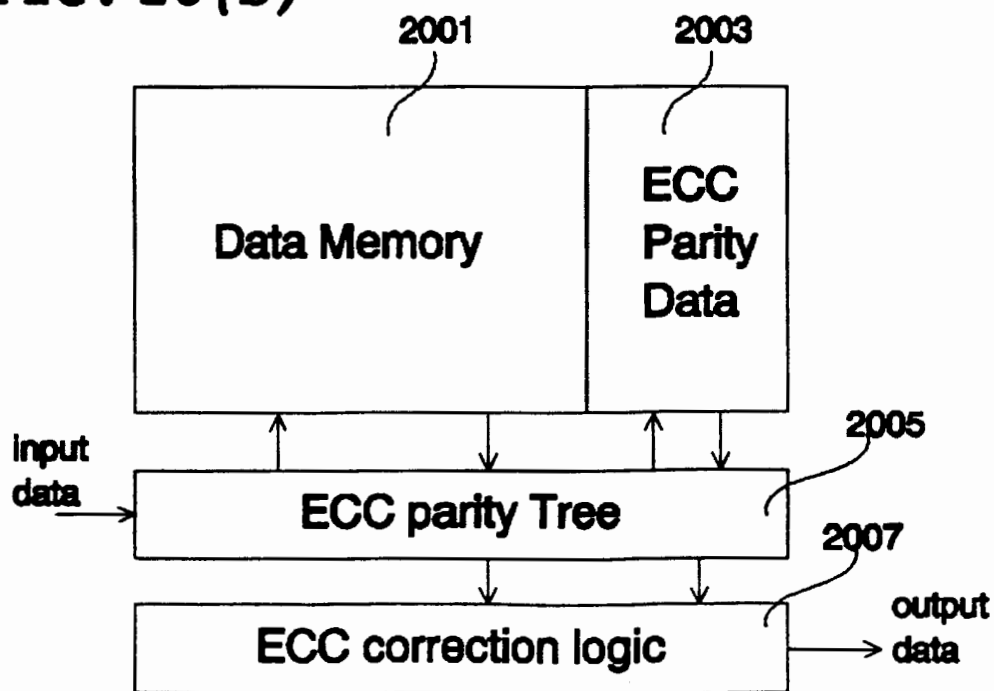




FIG. 19



**FIG. 20(a)****FIG. 20(b)**



6,108,229

1

# **HIGH PERFORMANCE EMBEDDED SEMICONDUCTOR MEMORY DEVICE WITH MULTIPLE DIMENSION FIRST- LEVEL BIT-LINES**

This is a Continuous-In-Part (CIP) Application of a previously filed application with Ser. No. 08/653,620, U.S. Pat. No. 5,748,547 filed on May 24, 1996 and another application Ser. No. 08/805,290 U.S. Pat. No. 5,825,704 filed on Feb. 25, 1997 by identical sole inventor as for this CIP Application.

## **BACKGROUND OF THE INVENTION**

### **1. Field of the Invention**

The present invention relates to high performance semiconductor memory devices, and more particularly to embedded memory devices having first level bit lines connected along different layout directions.

### **2. Description of the Prior Art**

DRAM is usually considered as a high density, low cost, but low performance memory device. DRAM's of current art always have lower performance relative to other types of semiconductor memories such as static random access memory (SRAM). The density of DRAM has been improved rapidly; the extent of integration has been more than doubled for every generation. Such higher integration of DRAM has been realized mainly by super fine processing technique and improvements in memory cell structure. In the mean time, the improvement in DRAM performance is progressing at a much slower rate. This relatively slower improvement rate in performance generates a performance gap between logic devices and memory devices. Many new approaches have been proposed to reduce this performance gap. The synchronized DRAM (SDRAM), the extended data output (EDO) DRAM, the multiple bank DRAM (MDRAM), and the RAMBUS system approaches are the most well known methods to improve DRAM performance. U.S. Pat. No. 4,833,653 issued to Mashiko et al. and U.S. Pat. No. 4,758,993 issued to Takemae et al. disclosed DRAM having selectively activated subarrays in order to improve performance. Another approach to improve DRAM performance is to place an SRAM cache into DRAM (called "hybrid memory"). U.S. Pat. No. 5,421,000 issued to Fortino et al., U.S. Pat. No. 5,226,147 issued to Fujishima et al., U.S. Pat. No. 5,305,280 issued to Hayano et al. disclosed embodiments of hybrid memories. The major problem for above approaches is that they are paying very high price for performance improvement, while the resulting memory performance improvement is still not enough to fill the gap. Another problem is that all of those approaches require special system design that is not compatible with existing computer systems; it is therefore more difficult to use them in existing computer systems.

Another disadvantage of DRAM is the need to refresh its memory. That is, the users need to read the content of memory cells and write the data back every now and then. The system support for DRAM is more complex than SRAM because of this memory refresh requirement. Memory refresh also represents a waste in power. U.S. Pat. No. 5,276,843 issued to Tillinghast et al. disclose a method to reduce the frequency of refresh cycles. U.S. Pat. No. 5,305,280 issued to Hayano et al. and U.S. Pat. No. 5,365,487 issued to Patel et al. disclosed DRAM's with self-refresh capability. Those inventions partially reduce power consumption by refresh operations, but the magnitude of power saving is very far from what we can achieve by the

2

present invention. The resource conflict problem between refresh and normal memory operations also remains unsolved by those patents.

Recently, Integrated Device Technology (IDT) announced that the company can make DRAM close to SRAM performance by cutting DRAM into small sub-arrays. The new device is not compatible with existing memory; it requires special system supports to handle conflicts between memory read operation and memories refresh operation. It requires 30% more area the DRAM, and its performance is still worse than SRAM of the same size.

Another important problem for DRAM design is the tight pitch layout problem of its peripheral circuits. In the course of the rapid improvement in reducing the size of memory cells, there has been no substantial improvement or change as to peripheral circuits. Peripheral circuits such as sense amplifiers, decoders, and precharge circuits are depend upon memory cell pitch. When the memory cells are smaller for every new generation of technology, it is more and more difficult to "squeeze" peripheral circuits into small pitch of memory layout. This problem has been magnified when the memory array is cut into smaller sub-arrays to improve performance. Each subarray requires its own peripheral circuits; the area occupied by peripheral circuits increases significantly. Therefore, in the foreseeable future, there may occur a case wherein the extent of integration of DRAM is defined by peripheral circuits. U.S. Pat. No. 4,920,517 issued to Yamauchi et al. disclosed a method to double the layout pitch by placing sense amplifiers to both ends of the memory. This method requires additional sense amplifiers. Although the available layout pitch is wider than conventional DRAM, the layout pitch is still very small using Yamauchi's approach.

All of the above inventions and developments provided partial solutions to memory design problems, but they also introduced new problems. It is therefore highly desirable to provide solutions that can improve memory performance without significant degradation in other properties such as area and user-friendly system support.

Another difficulty encountered by those of ordinary skill in the art is a limitation that Dynamic Random Access Memory (DRAM) which is usually considered as a high density, low cost, and low performance memory device cannot be conveniently integrated as embedded memory. This is due to the fact that higher integration of DRAM has been realized mainly by super fine processing technique and improvements in memory cell structure. A typical DRAM manufacture technology of current art is the four layer poly silicon, double layer metal (4P2M) process. Such memory technology emphasizes on super-fine structure in manufacture memory cells; performance of it logic circuit is considered less important. A technology optimized to manufacture high speed logic products have completely different priority; it emphasizes on performance of transistors, and properties of multiple layer metals. An example of a typical logic technology of current art is the triple layer metal, single poly silicon (1P3M) technology.

An embedded memory, by definition, is a high density memory device placed on the same chip as high performance logic circuits. The major challenge to manufacture high density embedded memory is the difficulty in integrating two types of contradicting manufacture technologies together. An embedded technology of current art requires 4 layers of poly silicon and 3 layers of metal. There are more than 20 masking steps required for such technology. It is extremely difficult to have reasonable yield and reliability

6,108,229

3

from such complex technology of current art. Further more, the current art embedded technology tend to have poor performance due to contradicting requirements between logic circuits and memory devices. None of current art embedded memory technology is proven successful. There is an urgent need in the Integrated Circuit (IC) industry to develop successful embedded memory devices.

The Applicant of this Patent Application has been successful in manufacturing embedded memory devices by novel approaches to change the architecture of IC memory so that the memory device no longer has conflicting properties with logic circuits. Examples of such architecture change have been disclosed in co-pending patent application Ser. No. 08/653,620. The previous application solved the tight pitch layout problems along the sense amplifier location, and it solves the self-refresh requirement by hiding refresh requirements. This CIP Application further discloses solutions for remaining problems. A single-transistor decoder circuit solves the tight pitch layout problem along the decoder direction. Typical logic technology or small modification of existing logic technology may be applied to manufacture the memory cells. Using these novel inventions, high performance and high density embedded memory devices are ready to be manufactured.

#### SUMMARY OF THE PRESENT INVENTION

The primary objective of this invention is, therefore, to improve the performance of semiconductor memory device without paying extensive area penalty. Another primary objective is to make DRAM more user-friendly by making the performance improvement in parallel with simplification in system supports. Another primary objective is to provide an improved semiconductor memory device in which peripheral circuits can readily follow further higher integration of memory cells. Another objective is to reduce power consumption of high performance semiconductor memory.

Another important objective of this invention is to manufacture high-density memory device on the same chip with high performance logic devices without using complex manufacture technology. Another primary objective is to make embedded DRAM to have the same performance as high-speed logic circuits. Another primary objective is to improve yield and reliability of embedded memory products.

These and other objects are accomplished by a semiconductor memory device according to the invention. The memory device includes a novel architecture in connecting bit lines along multiple layout directions, a new design in decoder circuit, and a novel timing control that can finish a read cycle without waiting for completion of memory refresh.

According to the present invention as described herein, the following benefits, among others, are obtained.

(1) The multiple dimensional bit line structure dramatically reduces the parasitic loading of bit lines seen by sense amplifiers. Therefore, we can achieve significant performance improvement. Our results show that a memory of the present invention is faster than an SRAM of the same memory capacity.

(2) The multiple dimension bit line structure also allows us to use one sense amplifier to support many bit line pairs. Therefore, we no longer have tight pitch layout problem for sense amplifiers and other peripheral circuits. Removing tight pitch problem allows us to achieve performance improvement without paying high price in layout area.

(3) A novel decoder design reduces the size of memory decoder dramatically, that allow designers to divide the

4

memory array into sub-arrays without paying high price in the area occupied by decoders.

(4) A novel input and output (IO) circuit design allows us to delay the memory refresh procedures until next memory operation. This approach allows us to "hide" refresh cycles and memory update cycles in a normal memory operation. The resulting memory device is as friendly as existing SRAM device. In fact, a memory of this invention can be made fully compatible with existing SRAM device.

(5) All of the above improvements are achieved by using much lower power than the power used by prior art DRAM's.

(6) The tight pitch layout problem along the decoder direction is solved. Therefore, we can divide a memory array into smaller blocks without sacrificing significant area. This architecture change allows us to use smaller storage capacitor for each DRAM memory cell, which simplifies manufacture procedure significantly.

(7) High density DRAM memory cells can be manufacture by adding simple processing steps to logic IC technology of current art. The resulting product supports high performance operation for both the memory devices and the logic circuits on the same chip.

(8) The simplification in manufacturing process results in significant improvements in product reliability and cost efficiency.

While the novel features of the invention are set forth with particularly in the appended claims, the invention, both as to organization and content, will be better understood and appreciated, along with other objects and features thereof, from the following detailed description taken in conjunction with the drawing.

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a schematic block diagram of a prior art memory device;

FIG. 2 is a simplified block diagram of a multiple bank semiconductor memory device;

FIG. 3a is a schematic block diagram of a memory device with two-dimensional bit lines;

FIG. 3b is a schematic block diagram of a memory device with three-dimensional bit lines;

FIG. 4a is an illustration showing layout and power consumption of a prior art memory bank;

FIG. 4b is an illustration showing layout and power consumption of a semiconductor memory device of a first embodiment according to the invention;

FIG. 5 is a schematic diagram of the sense amplifier used by this invention;

FIG. 6 is a schematic diagram of the IO circuits of the present invention;

FIG. 7a shows the waveforms of critical signals during a read cycle;

FIG. 7b shows the waveforms of critical signals during a refresh cycle;

FIG. 7c shows the waveforms of critical signals during a write cycle;

FIG. 8 is a schematic diagram of the IO circuits of the present invention to support faster data read; and

FIG. 9 shows the timing relationship of critical signals of a memory device of this invention.

FIG. 10 shows an example of a prior art CMOS decoder;

FIG. 11(a) is a schematic diagram of an enhance mode single transistor decoder of the present invention, and FIG.



6,108,229

5

11(b) is a diagram for the control signals and output signals of the decoder in FIG. 11(a);

FIG. 12(a) is a schematic diagram of a depletion mode single transistor decoder of the present invention, and FIGS. 12(a,b) illustrate the control signals and output signals of the decoder in FIG. 12(a);

FIG. 13 is a schematic diagram of a memory cell that uses an active transistor device as the storage capacitor of the memory cell;

FIGS. 14(a-g) are cross-section diagrams describing the process step to manufacture a DRAM memory cell by adding one masking step to standard logic technology;

FIGS. 15(a-c) are top views of the process step to manufacture a DRAM memory cell by adding one masking step to standard logic technology;

FIGS. 16(a-d) are cross-section diagrams describing another process step to manufacture a self-aligned trench capacitor for DRAM memory cell using one additional mask to standard logic technology;

FIG. 17 shows the top view of the memory cell manufactured by the process illustrated in FIGS. 17(a)-(d);

FIG. 18(a) shows the cross-section structures for capacitors that do not have the electrode voltage polarity constraint;

FIG. 18(b) shows the cross-section structures for memory cells that use transistors to separate nearby trench capacitors;

FIG. 19 illustrates the top view structure of practical memory cells of the present invention;

FIG. 20(a) shows a typical distribution of memory refresh time for the memory cells in a large DRAM; and

FIG. 20(b) is a symbolic diagram for a DRAM equipped with error-correction-code (ECC) protection circuit

#### DETAILED DESCRIPTION OF THE INVENTION

Before the invention itself is explained, a prior art semiconductor memory-device is first explained to facilitate the understanding of the invention.

FIG. 1 shows memory cell array structure of a prior art DRAM in both electrical and topographical manners. Memory cell array 100 includes plural pairs of bit lines BL1, BL1#, BL2, BL2#, BL3, BL3#, . . . ; BLn, BLn# (n; integer) which are disposed in parallel manner and a plurality of word lines WL1, WL2 . . . WLn (m; integer) which are disposed in parallel manner and also in such manner that they intersect with bit lines perpendicularly. At intersecting points, memory cells MC1, MC2, . . . , MCn are disposed. Memory cells are shown by circle marks in memory cell array 100 in FIG. 1. Each memory cell includes a switching field effect transistor 110 and memory cell capacitor 112. Bit line BL is connected to the drain of the transistor 110. The gate of transistor 110 is connected to word line WL. Sense amplifiers SA1, SA2, . . . SAn are disposed at one end of memory cell array and each pair of bit lines are connected to one sense amplifier. For example, a pair of bit lines BL1, BL1# are connected to sense amplifier SA1, a pair of bit lines BL2, BL2# are connected to sense amplifier SA2 . . . , and a pair of bit lines BLn, BLn# are connected to sense amplifier SAn. The outputs of those sense amplifiers are connected to data output switches 120. The output switches 120 contain a multiplexer 122 that is controlled by a decoder 124. The output switches 120 select the outputs from one of the sense amplifiers, and place the data on the data buses D and D#.

For example, when information is read out from memory cell MC1, the following operations are carried out. First,

6

word line WL2 is selected by the word line decoder 126 and the transistor 110 in memory cell MC1 is rendered conductive. Thereby, signal charge in capacitor 112 of memory cell MC1 is read out to bit line BL1# so that minute difference of electric potential occurs between a pair of bit lines BL1 and BL1#. The sense amplifier SA1 amplifies such difference. The output switches 120 select the outputs of SA1 and thereafter, transfer the data to data buses D, D# through a multiplexer 122. After the above read procedure, the charge stored in the cell capacitor 112 is neutralized. It is therefore necessary to write the original data sensed by SA1 back to the memory cell MC1. Such procedure is called "refresh". The sense amplifier used in current art always refreshes the memory cell after it determines the state of the memory cell. It is very important to remember that all the other memory cells along the word line, MC2, MC3, . . . MCn, are also rendered conductive when WL2 is selected. It is therefore necessary to turn on all the other sense amplifiers SA2, SA3, . . . SAn to read and refresh the data stored in all other memory cells connected to WL2, when we only need the data stored in MC1.

DRAM of such structure has the following drawbacks.

(1) In order to read the data from a few memory cells along one word line, we need to read and refresh all the memory cells along that word line. Most of the energy is used for refreshing instead of reading data. This waste in energy also results in slower speed because a large number of devices need to be activated.

(2) As the size of the memory array increases, the bit line parasitic capacitance (Cb) increases. The ratio between the memory cell capacitance Cm and the bit line parasitic capacitance Cb determines the amplitude of the potential difference on the bit line pairs. The memory read operation is not reliable if the (Cm/Cb) ratio is too small. Thereby, the (Cm/Cb) ratio is often the limiting factor to determine the maximum size of a memory array. Special manufacturing technologies, such as the trench technology or the 4-layer poly technology, have been developed to improve the memory cell capacitance Cm. However, the Cm/Cb ratio remains a major memory design problem.

(3) To support refresh procedures, we always need to have one sense amplifier for each bit line pair. As higher integration of memory cells progresses, the layout pitch for sense amplifier decreases. Thereby, it becomes difficult to form stable and well operable sense amplifier within the pitch. Such problem is often referred as the "tight pitch layout" problem in the art of integrated circuit design. Tight pitch layout always results in excessive waste in silicon area due to the difficulty in squeezing active devices into a narrow space. Similar problem applies to other peripheral circuits such as decoders and pre-charge circuits.

To reduce the effect of the above problems, large memory of prior art is always divided into plural sub-arrays called memory banks 200 as shown in FIG. 2. Each bank 200 of the memory sub-array has its own decoder 210 and output switches 212. Each pair of the bit lines in each memory bank needs to have one sense amplifier 214. The outputs of each memory bank are selected by output switches 212, and placed on data buses 220 so that higher order amplifiers and decoders can bring the data to output pins.

This multi-bank approach provides partial solutions to the problems. Because each memory bank is capable of independent operation, we can reduce power consumption by keeping unused memory banks in low power state. The speed is also improved due to smaller active area. The (Cm/Cb) ratio can be kept at proper value by limiting the

6,108,229

7

size of each memory bank. Multiple-bank memory allows us to turn on a sub-set of sense amplifiers to save power, but each bit line pair still needs to have one sense amplifier because we still need to refresh the contents of all activated memory cells. This multi-bank approach provides partial solutions, but it creates new problems. Each memory bank needs to have a full set of peripheral circuits; the areas occupied by the peripheral circuits increase significantly. Smaller size of memory bank implies higher percentage of area spent on peripheral circuits. Balancing the requirement between (Cm/Cb) ratio and the increase in tight pitch layout peripheral circuits is a major design problem for multiple bank memories. Yamauchi et al. were able to double the pitch for sense amplifiers by placing sense amplifiers at both sides of the memory array, but the layout pitch is still too small. Many other approaches have been proposed, but all of them provided partial solutions to part of the problems while created new problems.

This invention is made to solve the above-stated problems. FIG. 3a shows memory structure of one embodiment of the present invention in both electrical and topographical manners. The building block of the present invention is a memory unit 300. Each memory unit contains decoders 302, amplifiers AMP1, AMP2, . . . , AMPi, and a plurality of memory blocks 310. These memory blocks are arranged in pairs; memory block 1# is symmetrical to memory block 1; memory block 2# is symmetrical to memory block 2; . . . ; and memory block i# is symmetrical to memory block i. Each memory block contains word line switches 312, bit line switches 314, and a small memory array 316. The word line switches 312 and bit line switches 314 are controlled by block select signals. For example, the block select signal BLKSEL1 controls the word line switches and the bit line switches in memory block 1 and in memory block 1#. The memory array contains memory cells similar to the memory cells in FIG. 1. Circle marks are used to represent those memory cells in FIG. 3a. Each memory cell is connected to a short word line and a short bit line within each memory block. For example, in memory block 1 the gate of the memory cell MC12 is connected to block word line WL12 and block bit line BL12. Each block word line is connected to one unit word line through a word line switch 312. For example, WL12 is connected to UWL2 through a word line switch 312 controlled by block select signal BLKSEL1; WL22 is connected to UWL2 through a word line switch controlled by block select signal BLKSEL2; . . . ; WLij is connected to UWLj through a word line switch controlled by block select BLKSELi (i and j are integers). In this example, the memory unit has two levels of bit lines—the unit level bit lines UBL1, UBL1#, UBL2, BL2# . . . UBLn, UBLn# and the block level bit lines BL11, BL11#, BL12, BL12#, . . . et al. The block bit lines are made by the first layer metal (metal 1), and they are disposed vertical to the word lines. The unit bit lines are made by the second layer metal (metal 2), and they are disposed in parallel to the word lines. Each block bit line is connected to one unit bit line through one bit line switch 314 in each block. For example, BL12 is connected to UBL2 through a bit line switch controlled by block select signal BLKSEL1; BL22 is connected to UBL2 through a bit line switch also controlled by block select signal BLKSEL2; . . . ; BLii is connected to UBLi through a bit line switch controlled by block select BLKSELi. Each pair of unit bit lines is connected to one amplifier. For example, UBL1 and UBL1# are connected to AMP1; UBL2 and UBL2# are connected to AMP2; . . . ; UBLi and UBLi# are connected to AMPi. Those unit-bit-lines and block-bit-lines form a two-dimensional network that allows one amplifier to support bit line pairs in many blocks.

8

This two-dimensional bit line connection allows us to read the memory content with little waste in power. For example, when information is read out from memory cells on WL12 in block 1, the following operations are carried out. First, the block-select signal BLKSEL1 is activated, while all other block select signals remain inactive. All the word line switches 312 and bit line switches 314 in memory block 1 and in memory block 1# are rendered conductive, while those of all other memory blocks remain inactive. The unit decoder 302 activates the unit word line UWL2 while keeping other unit word lines inactive. Therefore, only WL12 is activated while all other block word lines remain inactive. The transistor 110 in memory cell MC12 is rendered conductive. Thereby, signal charge in capacitor of memory cell MC12 is read out to block bit line BL12 and to unit bit line UBL2 through the block bit line switch 314. In the mean time, BL12# is also connected to UBL2# through the block bit line switch in memory block 1#, but there is no signal charge read out to UBL2# because WL12# remains inactive. Since the bit lines in the memory block pairs are drawn in mirror symmetry, their parasitic capacitance is matched. The signal charge in memory cell MC12 develops a minute difference of electric potential between UBL2 and UBL2#. Such difference is detected and is amplified by sense amplifier AMP2; the result is sent to high order data bus (not shown), and is used to refresh memory cell MC12. Similarly, the content of memory cell MC11 is read and refreshed by sense amplifier AMP1; the content of memory cell MCi1 is read and refreshed by sense amplifier AMPi.

If we want to read the data from memory cells on WL12# in block 1#, the procedure is identical except that the unit decoder 302 should activate UWL2# instead of UWL2. If we want to read from memory cells in WLij in block i, the unit decoder 302 should turn on UWLj and the block select signal BLKSELi should be activated. The content of memory cell MCi1 is read and refreshed by sense amplifier AMP1; the content of memory cell MCi2 is read and refreshed by sense amplifier AMP2; . . . ; and the content of memory cell MCii is read and refreshed by sense amplifier AMPi.

It is still true that one sense amplifier is activated for each activated memory cell; otherwise the data stored in the memory cell will be lost. The differences are that the activated sense amplifiers no longer need to be placed right next to the local bit lines connected to the activated memory cell and that the number of activated memory cells is only a small fraction of that of a prior art DRAM. The multiple dimensional bit line structure allows us to place the activated sense amplifier far away from the activated memory cells without introducing excessive parasitic loading to the bit lines. The layout pitches of sense amplifier and peripheral circuits are independent of the size of memory cell. It is therefore possible to design high performance peripheral circuits without increasing the area significantly.

It is to be understood that the present invention describes multiple dimension bit line structure “before” the first level sense amplifiers detect the storage charges in the activated memory cells. Prior art multi-bank DRAM often has multiple dimension data buses “after” the first level sense amplifier already detected the storage charge in activated memory cells. The prior art multi-bank memories need one first level sense amplifier for every bit line pairs, and they do not solve the tight pitch layout problem.

While specific embodiments of the invention have been illustrated and described herein, it is realized that other modification and changes will occur to those skilled in the art. For example, the above embodiment assumes that bit

6,108,229

9

line pairs are rendered in opposite memory block pairs. It should be obvious to those skilled in the art that this invention also can support the conventional bit line pairing structure in FIG. 1 where bit line pairs are arranged right next to each other. It is also obvious that the above two-dimensional bit line structure can be easily expanded to three-dimensional or multi-dimensional bit line structures. A two dimensional bit line structure is described in FIG. 3a for its simplicity, but the number of levels of bit line structures is not limited by the above example. The optimum levels of bit line structures are determined by details of manufacture technology and by the design specifications.

It also should be obvious that the bit line switches are not required elements; the unit bit lines can be connected directly to block bit lines without bit lines switches. Bit line switches help to reduce the bit line capacitance seen by each sense amplifier, but they are not required for functional reason because the word line switches already can isolate the memory cells in each memory block from memory cells in other memory blocks. While one sense amplifier is placed in each pair of memory block in the above example, there is no such constraint in this invention. We can place more than one sense amplifier per memory block, or place one sense amplifier in the area of many memory blocks. Because of a structure of multiple dimension bit line, the present invention completely removes the layout constraint between memory array and peripheral circuits.

FIG. 3b shows a memory array of the present invention with 3-level bit line connections. For simplicity, only two pairs of bit lines are shown in this figure. The first level of bit lines are made by the first layer metal (M1), the second level is made by the second layer metal (M2), and the third level is made by the third layer metal (M3). Each memory block 350 contains a plurality of side-by-side M1 bit line pairs (BBLi, BBLi#), (BBLj, BBLj#). This memory array contains a plurality of memory columns 360. The M1 bit lines are connected to corresponding M1 bit lines in other memory blocks along the same memory column 360 by M2 bit lines CBLi, CBLi#, CBLj, CBLj#. The bit lines in each column are connected to the bit lines in other columns using metal 3 bit lines M3Li, M3Li#, M3Lj, M3Lj# through bit line switches 362. For each bit line in one memory column 360, we only need one bit line switch 362 and one M3 bit line. A group of sense amplifiers SA1, . . . , Sai, . . . SAj, are placed at one end of the memory array. Each pair of the above three-dimension bit line networks are connected to one sense amplifier. For example, the (BBLi, CBLi, M3Li), (BBLi#, CBLi#, M3Li#) pair are connected to SAi, and the (BBLi, CBLi, M3Li), (BBLi#, CBLi#, M3Li#) pair are connected to SAj. Since each memory block 350 has its own word line switch (not shown in FIG. 3b), no more than one memory block in the network can be activated at any time. It is therefore possible to support a large number of memory cells using a small number of sense amplifiers without violating the requirement that every activated memory cell must have an activated sense amplifier to detect its storage charge.

Although the bit line structure in FIG. 3b is the actual bit line structure used in our product, for simplicity, we will use the simpler two-dimensional bit line structure in FIG. 3a as example in the following discussions.

The difference in layout area and the difference in power consumption between the prior art and this invention are illustrated by the simplified block diagrams in FIGS. 4(a,b). FIG. 4a shows a simplified symbolic graph of one memory bank of conventional DRAM memory array 400 that has N bit line pairs, M word lines, and 8 output (N and M are

10

integers). The sense amplifiers are represented by long rectangles 402 in FIG. 4a. Because one sense amplifier supports each bit line pair, the layout pitch for the sense amplifier is the layout pitch of a bit line pair, so that they must be placed in long narrow rectangular area. The outputs of the sense amplifiers are selected into 8 outputs by the output decoder 404 and multiplexers 406. The layout pitch for the output decoder 404 is also very narrow. The layout pitch for each element of the word line decoder 410 is the pitch of one memory cell Cx. For a memory operation, one word line 412 is activated across the whole memory bank. The number of active memory transistors is N. All N sense amplifiers are activated, and all N bit line pairs in this memory bank are charged or discharged by the sense amplifiers. The activated area covers the whole memory bank as illustrated by the shaded area in FIG. 4a.

FIG. 4b is a simplified symbolic graph of one bank of DRAM memory array of the present invention. For simplicity in comparison, we assume that the memory array in FIG. 4b contains the same number of memory cells and the same number of data outputs as the memory array in FIG. 4a. The memory bank is divided into 4 units 450, and each unit contains 8 pairs of memory blocks 452. We have one amplifier 454 for each pairs of memory blocks. Each unit has one unit word line decoder 456. Detailed structure of the memory unit has been described in FIG. 3a. A unit select decoder 460 generates unit select signals XBLKSEL along word line directions. A block select decoder 462 generates bank level block select signals YBLKSEL. A memory block 452 is activated when both XBLKSEL and YBLKSEL crossing the block are activated. The local block select signals are generated by AND gates in the amplifier 454 area. The outputs of each amplifier is placed on bank level bit lines KBL, KBL# to input/out (IO) units 470 at the edge of the memory. For simplicity, only one pair of bank level bit lines are shown in FIG. 4b. Further details of those peripheral circuits will be discussed in following sections. FIG. 4b shows that the layout pitch for the sense amplifiers 454 is 8 times wider than that in FIG. 4a. The peripheral circuits no longer require tight pitch layout, so that we can design them efficiently for both speed and area consideration. For a memory operation, only one memory block 452 and 8 sense amplifiers 454 in the selected unit 450 are activated. The shaded area in FIG. 4b illustrates the activated area. This active area is obviously much smaller than the active area of a conventional memory bank shown in FIG. 4a. Power consumption of the present invention is therefore much less than that of a prior art memory.

The parasitic bit line parasitic capacitance Cbp of the prior art memory in FIG. 4a is

$$C_{bp} = (M/2) * C_d + M * C_{m1} \quad (1)$$

And, where Cd is the diffusion capacitance for one bit line contact, Cm1 is the metal 1 capacitance of the bit line for each unit cell, and M is the number of memory cells along one bit line. We assume that two memory cells share each contact so that the total number of contacts is M/2.

The parasitic bit line capacitance Cb of the memory in FIG. 4b is

$$C_b = (M/16) * C_d + (M/8) * C_{m1} + (8 * C_d + N * C_{m2}) \quad (2)$$

where Cm2 is the metal 2 bit line capacitance for each memory pitch along the unit bit line direction. The first two terms (M/16)\*Cd+(M/8)\*Cm1 are the capacitance for a



6,108,229

11

local bit line that is  $\frac{1}{8}$  of the length of the bit line in FIG. 4a. The last two terms ( $8 \cdot C_d + N \cdot C_{m2}$ ) are the parasitic capacitance of the unit bit line that has 8 contacts to the bit line switches and a metal 2 bit line. The contact capacitance  $C_d$  is much larger than the metal capacitance. The metal 2 capacitance  $C_{m2}$  is usually smaller than the metal 1 capacitance  $C_{m1}$ . Therefore, Eqs. (1,2) show that the bit line parasitic capacitance seen by one sense amplifier of the present invention,  $C_b$ , is significantly smaller than  $C_{bp}$ . Smaller bit line capacitance implies faster speed, lower power, and better reliability. There is no need to use complex technology to build the memory cells. It is also possible to increase the size of each memory block to connect more memory cells to each sense amplifier in order to reduce the total area.

The total areas occupied by memory cells are identical between the two memory arrays in FIG. 4a and FIG. 4b. Therefore, the difference in area is completely determined by the layout of peripheral circuits. The available layout pitch for sense amplifiers and for output decoders for the memory in FIG. 4b is 8 times larger than that of the memory in FIG. 4a. It should be obvious to those skilled in the art that a memory of the present invention is smaller than a prior art memory along the dimension vertical to the word line direction due to wider layout pitch. Along the dimension in parallel to word lines, the present invention still needs a decoder 460 of the same layout pitch. In addition, this invention needs to have one set of word line switches 462 for each memory block 452. The additional area occupied by the word line switches 462 does not increase the layout area significantly because we can use smaller high level decoders due to reduction in loading.

The sense amplifier used in the present invention is substantially the same as typical sense amplifiers used in the prior art FIG. 5 shows schematic diagram of the amplifier in FIG. 3a. When the sense amplifier enable signal SAEN is activated, transistors MP11, MP12, MN11, and MN12 form a small signal sensing circuit that can detect minute potential difference on the unit bit line pairs UBL and UBL#. The transfer gate transistor MN14 transfers the signal between the unit level bit line UBL and the bank level bit line KBL when the bank level word line KWL is active. The transfer gate transistor MN13 transfers the signal between the unit level bit line UBL# and the bank level bit line KBL# when the bank level word line KWL is active. MN17 is used to equalize the voltages on UBL and UBL# when the sense amplifier is not active. The operation principles of the above sense amplifiers are well known to the art of memory design so we do not describe them in further details.

FIG. 6 is a block diagram of the IO unit 470 in FIG. 4b. The bank level bit line pair KBL and KBL# are connected to a bank level sense amplifier 650 through a bank level bit line switch 651. This sense amplifier 650 is identical to the sense amplifier in FIG. 5; its enable signal is KSAEN. The KBL switch 651 is rendered conductive when its enable signal MREAD is active, and it isolates the bit lines from the sense amplifier when MREAD is not active. This bit line switch 651 is used to improve the speed of the sense amplifier as well known to the art of memory design. The output of the sense amplifier, SOUT, is connected to an Error-Correction-Code (ECC) circuit 652. The ECC circuit is well known to the art, so we do not discuss it in further details. The output of the ECC circuit, EOUT, is connected to the input of an output driver 665. The output driver 665 drives the data to external pad when it is enabled by the signal READOUT. For a write operation, we place the data on the pad into a storage register 662. The output of the

12

storage register, UDATA, is connected to a memory write driver 664. The memory write driver 664 is controlled by the UPDATE signal to drive data on KBL and KBL# during a memory update operation.

FIGS. 7(a-c) show the waveforms of critical signals for the memory described in previous sections.

FIG. 7a shows the timing of critical signals during a memory operation to read data from memory cells (called a "read cycle"). First, the block select signal BLKSEL is activated at time T1. BLKSEL is active when both XBLKSEL and YBLKSEL are active. Whenever BLKSEL is active, the precharge circuit of the selected memory block is turned off, so does the precharge circuit of all the sense amplifiers of the selected memory unit. The precharge signal and bank level block select signals XBLKSEL, YBLKSEL are not shown in waveforms because the information is redundant with respect to BLKSEL signal. After BLKSEL is active, block word line WL is active at time T2. Once WL is active, a minute potential difference starts to develop in block bit line pair BL, BL# as well as unit bit line pair UBL, UBL#. After enough potential difference has developed on the unit bit line pairs, the sense amplifiers of the selected memory unit are activated by bring SAVCC to VCC, and SAVSS to VSS. The unit sense amplifier starts to magnify the bit line potential once it is activated at T3. The bank level word line KWL is then activated at T4; the potential differences in UBL and UBL# are transferred to bank bit line pairs KBL and KBL# once KWL is activated. Between time T4 and T5, the voltages of UBL and UBL# are first drawn toward PCGV due to charge sharing effect between bank bit lines and unit bit lines; the unit sense amplifier eventually will overcome the charge sharing effect and magnify their potential difference. At time T5, the bank-word-line KWL is off, and the pulling of KSAVCC to VCC and KSAVSS to VSS activates the bank level sense amplifier 750. The bank level sense amplifier 750 will magnify the potential difference on KBL and KBL# to full power supply voltages. In the mean time, the unit level sense amplifier will also pull UBL and UBL# to full power supply voltage. Because we are relying on the unit level sense amplifier to refresh the selected memory cells, we need to provide a timing margin to make sure the signal charges in those memory cells are fully restored before we can turn off the word line WL at T6. After the word line is off, sense amplifiers are deactivated at T7, then the block select signal BLKSEL is deactivated at T8. Once BLKSEL is off, the memory is set into precharge state, and all bit line voltages return to PCGV. A memory of this invention has much shorter precharge time than prior art memories due to much lower loading on each level of its bit lines. At time T9, all signals are fully restored to their precharge states, and the memory is ready for next memory operation.

FIG. 7b shows the timing of critical signals for a memory operation to refresh the data of memory cells (called a "refresh cycle"). A refresh cycle is very similar to a read cycle except that we do not need to bring the data to bank level. All these bank level signals, KWL, KSAVCC, KSAVSS, KBL, and KBL# remain inactive throughout a refresh cycle. At time T11, the block select signal BLKSEL is active, then the word line WL is activated at time T12. Potential differences start to develop in block level and unit level bit lines BL, BL#, UBL, and UBL#. At time T13 the sense amplifier is activated. The sense amplifier quickly magnify and drive the bit lines to full power supply voltages. When the charges in selected memory cells are fully restored, we can turn off the word line WL at T14, then turn off block select signal BLKSEL at T15. At time T16, all the

6,108,229

13

signals are restored into precharge states, and the memory is ready for next operation. Comparing FIG. 7b with FIG. 7a, it is obvious that the time need for a fresh cycle is shorter than the time for a read cycle because we do not need to drive KBL and KBL#.

FIG. 7c shows the timing of critical signals during a memory operation to write new data into memory cells (called a "write cycle"). At time T21, the block-select-signal BLKSEL and bank level word line KWL are activated. In the mean time, the new data is written into the bank level bit lines KBL and KBL#, then propagate into lower level bit lines UBL, UBL#, BL, and BL#. The memory write driver 764 has strong driving capability so that bit lines can be driven to desired values quickly. At time T22, the unit level sense amplifier is activated to assist the write operation. Once the charges in the memory cells are fully updated, the word lines WL and KWL are turned off at T23. Then, the block select signal BLKSEL are off at T24. At T25 the memory is fully restored to precharge state ready for next memory operation. Comparing FIG. 7c with FIG. 7a, it is obvious that the time needed to execute a write cycle is much shorter than the time needed to execute a read cycle because of the strong driving capability of the memory write driver 764.

As illustrated by FIG. 7a, the reason why read operation is slower than write or refresh operations is because the read operation cannot be finished until the unit level sense amplifiers fully restore the signal charges in the selected memory cells. From the point of view of an external user, the additional time required to refresh the memory does not influence the total latency for a memory read operation because the process to deliver data from bank level circuit to external pad is executed in parallel. The refresh time is therefore "hidden" from external users. The only time an external user can feel the effect of this additional refresh time is when a refresh cycle is scheduled at the same time as a read cycle is requested. The memory can not execute a refresh cycle in parallel to a read cycle at a different address, so one of the requests must wait. External control logic is therefore necessary to handle this resource conflict condition. For a memory with ECC support, data write operations always need to start with memory read operations, so the above problems also apply to memory write operations. In order to fully compatible with an SRAM, we must make internal memory refresh cycles completely invisible to external users. This is achieved by simple changes in IO circuit shown in FIG. 8, and change in timing control shown in FIG. 9.

The IO circuit in FIG. 8 is almost identical to the IO circuit in FIG. 6 except that it has two additional multiplexers 854, 860. The output of the ECC circuit, EOUT, is connected to the input of a bypass multiplexer 854. During a read cycle, the bypass multiplexer 854 selects the output from the storage register 662 if the reading memory address matches the address of the data stored in the storage register 662. Otherwise, the bypass multiplexer 854 selects the output of the ECC circuit, and sends the memory output to the output driver 665. The storage multiplexer 860 selects the input from external pad during a write operation, and it selects the data from memory read out during a read operation. This architecture allows us to "hide" a refresh cycle in parallel with a normal memory operation. It also improves the speed of normal read operations. Using the circuit in FIG. 8, the most updated data of previous memory operation are always stored into the storage register 662. To execute a new memory operation, we always check if the data are stored in the storage register before reading data from the

14

memory array. If the wanted data is already stored in the storage register, no memory operation will be executed, and the data is read from the storage register directly. When a new set of data is read from the memory array, an update cycle is always executed before the end of a new memory operation to write the data currently in the storage buffer back into the memory array. Since we always store every memory read results into the storage registers, there is no need to refresh the selected memory cells immediately. With this configuration, we can terminate the read operation before the unit level sense amplifier can fully refresh the activated memory cells. Therefore, the unit level circuits are available for a refresh cycle at the same time when the memory is propagating the read data to the external pads. This architecture removes the conflict between refresh cycle and normal memory operations. The operation principle of this scheme is further illustrated by the waveforms in FIG. 9.

FIG. 9 shows the worst case situation when a memory operation overlaps with a refresh operation (to a different address or to the same memory block), and when there is a need to update data from the storage buffer at the same time. Under this worst case condition, the refresh cycle and the memory update cycle must be "hidden" in the memory read operation in order to avoid complexity in system support. On the other word, we must execute the refresh and update cycles in parallel without influencing the timing observable by an external user.

At time Tr1 in FIG. 9, the block select signal BLKSEL is activated for a read operation. At time Tr2, the word line WL is activated, then the unit sense amplifier is activated at Tr3. The unit level word line KWL is activated at Tr4, and the unit level sense amplifier is activated at time Tr5. Until time Tr5, the memory operations and waveforms are identical to those shown in the read cycle in FIG. 8a. The operation is different starting at Tr5; we are allowed to turn off the block select signal BLKSEL, the word lines WL, KWL, and the unit level sense amplifier simultaneously at Tr5 without waiting for full amplification of the memory data. The memory block quickly recovers to precharge state ready for next operation at time Tf1. During this time period, the unit level sense amplifier does not have enough time to fully amplify the signals in the lower level bit lines BL, BL#, UBL, and UBL#. Those activated memory cells no longer stores the original data. That is perfectly all right because the correct data will be stored in the storage register 662 in the following procedures. At time Tf1, the data are sensed by the bank level sense amplifier; the correct data will be remembered in the storage register 662 and updated into those selected memory in the next memory operation. Therefore, the data are not lost even when the storage charge in the memory cells are neutralized at this time. At the same time when we are waiting for the bank level circuits to propagate the new read data to external circuits, the unit level and block level memory circuits are available for a refresh operation. This hidden refresh cycle can happen at any memory address. The worst case timing happen when the refresh cycle happens at the same block that we just read. FIG. 9 shows the timing of the worst case condition. At time Tf1, BLKSEL is activated for the refresh cycle. A refresh cycle with identical waveforms as the waveforms in FIG. 8b is executed from time Tf1 to time Tf5. At time Tw1, the memory unit is ready for new operation, and the bank level read operation is completed. At this time, the IO unit 720 is executing ECC correction and the data is propagating to the pads. In the mean time, the bank level resources are available, so we take this chance to update the old data in the



storage register 762 back into the memory array by executing a write cycle. The waveforms in FIG. 9 from time Tw1 to Tw5 are identical to the waveforms in FIG. 7c. At the end of the memory operation, the latest data just read from the memory are stored into the storage register 662, the previous data are updated into the memory array, we fulfilled a refresh request, and the external memory operation request is completed.

It is still true that we need to record the data stored in every activated memory cell. Otherwise the data will be lost. The difference between the above memory access procedures and conventional DRAM memory accesses is that the data is temporarily stored in the storage registers so that we do not need to refresh the activated memory cells immediately. This architecture delays data update until next memory process using available bandwidth, so that refresh cycles and update cycles can be hidden to improve system performance.

The above architecture is different from a hybrid memory because (1) this invention simplifies the timing control of DRAM read cycle while the SRAM of the hybrid memory does not simplify the DRAM operation, (2) the system control and device performance of the present invention is the same no matter the memory operation hits the storage register or not, while the performance and control of a cache memory is significantly different when the memory operation miss the cache array, (3) a hybrid memory has better performance when the size of the SRAM cache is larger due to higher hit rate, while the performance of the present invention is independent of hit rate, and (4) the storage register does not introduce significant area penalty while the on-chip SRAM of hybrid memory occupies a significant layout area. The structure and the operation principles of the memory architecture described in the above sections are therefore completely different from the structures of hybrid memories.

As apparent from the foregoing, the following advantages may be obtained according to this invention.

(1) The tight pitch layout problem is solved completely. Since many bit line pairs share the same sense amplifier, the available layout pitch for each peripheral circuit is many times of the memory cell pitch. Therefore, sense amplifiers and peripheral circuits of high sensitivity with electrical symmetry and high layout efficiency can be realized.

(2) The bit line loading seen by the sense amplifier is reduced dramatically. It is therefore possible to improve the performance significantly.

(3) It is also possible to attach a large number of memory cells to each sense amplifier to reduce total device area.

(4) The novel design in decoder reduces decoder size significantly without sacrificing driving capability. The loading on each unit word line is also reduced significantly. This decoder design reduces layout area and improves device performance.

(5) Changes in memory access procedures allow us to delay the refresh operation until next memory operation. Internal refresh operations are therefore invisible for external users.

(6) The only devices activated in each memory operation are those devices must be activated. There is little waste in power. The present invention consumes much less power than prior art memories.

A memory device of the present invention is under production. Using 0.6 micron technology to build a memory array containing one million memory cells, we are able to achieve 4 ns access time, which is more than 10 times faster than existing memories devices of the same storage capacity.

FIG. 10 shows an example of a typical prior art decoder. Each branch of the decoder contains one AND gate 1101 that controls one of the outputs of the decoder O3-0. Two sets of mutually exclusive input select signals (G0, G0NN) and (G1, G1NN) are connected to the inputs of those AND gates as show in FIG. 10, so that no more than one output O3-0 of the decoder can be activated at any time.

FIG. 11(a) is the schematic diagram of a single-transistor decoder that uses only one n-channel transistor M3 to M0 for each branch of the decoder. The source of each transistor M3 to M0 is connected to one word line WL3 to WL0 of the memory array. A set of mutually exclusive drain select signals DSEL1, DSEL0 are connected to the drains of those transistors M3 to M0, and a set of mutually exclusive gate select signals GSEL1 and GSEL0 are connected to the gates of those transistors M3 to M0, as shown in FIG. 11(a). In this configuration, WL3 is activated only when both DSEL1 and GSEL1 are activated, WL2 is activated only when both DSEL1 and GSEL0 are activated, WL1 is activated only when both DSEL0 and GSEL1 are activated, and WL0 is activated only when both DSEL0 and GSEL0 are activated. Therefore, the circuit in FIG. 11(a) fulfills the necessary function of a memory word line decoder. A typical CMOS AND gate contains 3 p-channel transistors and 3 n-channel transistors. The decoder in FIG. 12(a) uses only one transistor for each output of the decoder. It is apparent that the decoder in FIG. 11(a) is by far smaller than the one in FIG. 10. However, the single-transistor decoder in FIG. 11(a) requires special timing controls as illustrated in the following example.

FIG. 11(b) illustrates the timing of input signals to activate one of the word line WL0. Before time T0, there are no decoding activities. All gate select signals GSEL1, GSEL0 must stay at power supply voltage Vcc, and all drain select signals DSEL1, DSEL0 must stay at ground voltage Vss. Otherwise one of the word line maybe activated accidentally by noise or leakage. To activate one word line WL0, we must deactivate all gate select signals GSEL1, GSEL0 at time T0, then activate one of the gate select signal GSEL0 and one of the drain select signal DSEL0 at T1. In order to deactivate the decoder, DSEL0 must be deactivated at time T2 before all gate select signals GSEL1 and GSEL0 are activated again at T3. The above control sequence is necessary to prevent accidental activation of word lines that are not selected. The above timing control sequence is complex because all inputs are involved when we only want to active one word line. The above decoders are simplified examples of 4 output decoders. A realistic memory decoder will need to control thousands of word lines. The power consumed by such complex control sequences can be significant for a realistic memory decoder. Another problem for the decoder in FIG. 11(a) is also illustrated in FIG. 11(b). Due to body effect of n-channel transistor M0, the voltage of the activated word line WL0 is lower than the power supply voltage Vcc by an amount Vbd as shown in FIG. 11(b). This voltage drop can be a big problem for a DRAM decoder because it will reduce the signal charge stored in DRAM memory cells.

FIG. 12(a) is a schematic diagram of a decoder of the present invention. The only differences between the decoders in FIGS. 11(a), 12(a) is that depletion mode transistors D3 to D0, instead of enhanced mode transistors M3 to M0, are used by the decoder shown in FIG. 12(a). The threshold voltage of those depletion mode transistors D3 to D0 is controlled to be around -0.2 volts (or roughly 1/3 of the threshold voltage of a typical enhance mode transistor) below power supply voltage Vss.

FIG. 12(b) illustrates the timing of input signals to select one word line WL0 of the depletion-mode single transistor

6,108,229

17

decoder in FIG. 12(a). Before time T<sub>0</sub>, all the gate select singles GSEL<sub>1</sub>, GSEL<sub>0</sub>, and all the drain select signals DSEL<sub>1</sub>, DSEL<sub>0</sub> are at ground voltage V<sub>ss</sub>. Unlike the enhance mode single transistor decoder in FIG. 11(a), it is all right to set the gate control signals GSEL<sub>1</sub>, GSEL<sub>0</sub> at V<sub>ss</sub> when the decoder is idle. The word lines WL<sub>3</sub>–WL<sub>0</sub> won't be activated by noise or small leakage because the depletion-mode transistors D<sub>3</sub> to D<sub>0</sub> are partially on when its gate voltage is at V<sub>ss</sub>. To activate one word line WL<sub>0</sub>, we no longer need to deactivate all gate select signals. We only need to activate one gate select signal GSEL<sub>0</sub> and one drain select signal DSEL<sub>0</sub> as shown in FIG. 12(b). To deactivate the decoder, we can simply deactivate GSEL<sub>0</sub> and DSEL<sub>0</sub> as shown in FIG. 12(b). This control sequence is apparently much simpler than the control sequence shown in FIG. 11(b). There is also no voltage drop cause by body effect on the selected word line because the threshold voltage of the activated transistor M<sub>0</sub> is below zero. The depletion mode single transistor decoder in FIG. 12(a) is equally small in area as the enhance mode single transistor decoder in FIG. 11(a), but it will consume much less power. The only problem is that some of those word lines are partially activated when they have deactivated gate select signal but activated drain select signal as illustrated by WL<sub>1</sub> in FIG. 12(b). This partial activation of word lines is not a functional problem when the voltage V<sub>pt</sub> is less than the threshold voltage of selection gates in the memory cells, but it may introduce a potential charge retention problem due to sub-threshold leakage current. One solution for this problem is to introduce a small negative voltage on all deactivated gate select signals at time T<sub>0</sub> as shown in FIG. 12(c). This small negative voltage V<sub>nt</sub> on the drain select signal assures the depletion gate transistor D<sub>1</sub> remains uncondutive so that the word line WL<sub>1</sub> won't be partially activated.

While specific embodiments of single transistor decoders have been illustrated and described herein, it is realized that other modifications and changes will occur to those skilled in the art. For example, p-channel transistors or depletion mode p-channel transistors can replace the n-channel transistors in the above examples.

As apparent from the foregoing, single-transistor-decoders of the present invention occupies much small area than the prior art CMOS-decoders. It is therefore possible to divide a large memory array into small block—each block isolated by its own decoder—without increasing the total area significantly. When the memory array is divided into small blocks, we no longer need to have large storage capacitor as prior art DRAM cells have. It is therefore possible to build DRAM memory cells using standard logic technology.

One example of DRAM memory cell built by logic technology is shown in FIG. 13. This memory cell 1400 contains one select transistor 1402 and one storage transistor 1404. The gate of the storage transistor 1404 is biased to full power supply voltage V<sub>cc</sub> so that it behaves as a capacitor. The drain of the storage transistor 1404 is connected to the source of the select transistor 1402. The gate of the select transistor 1402 is connected to word line WL, and the drain of the select transistor is connected to bit line BL. Using this memory cell 1400 and a memory architecture disclosed in this invention and in our previous patent application, commercial memory products were manufactured successfully. The major advantage of the logic memory cell 1400 is that it can be manufactured using standard logic technology. The resulting memory product achieved unprecedented high performance. The area of the logic memory cell 1400 is larger than prior art DRAM cells because two transistors,

18

instead of one transistor and one capacitor, are used to build one memory cell. It is therefore desirable to be able to build single transistor memory cell from a manufacture technology as similar to logic technology as possible.

Therefore, according to FIGS. 3a to 4b, and FIGS. 12(a) to 13, a semiconductor memory device 300 is disclosed which is provided for operation with a plurality of cell-refreshing sense-amplifiers (SAs). The memory device 300 includes a memory cell array having a plurality of first-direction first-level bit lines, e.g., bit line BL<sub>n</sub>i in block n for bit-i, along a first bit-line direction, disposed in a parallel manner along a first direction, e.g., a horizontal direction. The memory cell array further includes a plurality of word lines WL intersected with the first-direction first-level bit lines. The memory cell array further includes a plurality of memory cells. Each of these plurality of memory cells being coupled between one of the first-direction first level bit lines, i.e., bit line BL<sub>n</sub>i in block n for bit-i, along a first bit-line direction and one of the word lines for storing data therein. The memory device further includes a plurality of different-direction first level bit lines, e.g., multiple-block or the unit bit-line-i such as UBL<sub>i</sub>, BBL<sub>i</sub>, CBL<sub>i</sub>, etc. (referring to FIG. 3b), where i=1, 2, 3, . . . I, disposed along a plurality of different directions, e.g., along a vertical direction, with at least one of the different directions being different from the first direction, wherein each of the first direction first level bit lines connected to one of the cell-refreshing sense amplifiers (SAs) directly or via the different-direction first level bit-lines. In a specific preferred embodiment, one of the different directions, e.g., a vertical direction, for arranging the different-direction first level bit lines, e.g., the multiple-block bit-line-i UBL<sub>i</sub>, BBL<sub>i</sub>, CBL<sub>i</sub>, etc. (referring to FIG. 3b). Where i=1, 2, 3, . . . I, being perpendicular to the first direction, e.g., a horizontal direction for arranging the first-direction first level bit lines. In the preferred embodiment as shown in FIG. 4b, the memory device 300 further includes bit line switches connected between the first level bit lines, which are arranged in different directions. The semiconductor memory device further includes a decoder 302 for generating an activating signal for activating one of the word lines WL. The decoder 302 further includes a plurality of drain select lines, e.g., DSEL<sub>0</sub> AND DSEL<sub>1</sub>, etc., each being provided for receiving one of a plurality of mutual exclusively drain select signals. The decoder 302 further includes a plurality of gate select lines, e.g., GSEL<sub>0</sub>, GSEL<sub>1</sub>, etc., each being provided for receiving one of a plurality of mutual exclusively gate select signals. The decoder 302 further includes a plurality of transistors, e.g., D<sub>0</sub>, D<sub>1</sub>, or M<sub>0</sub>, M<sub>1</sub>, etc. Each transistor includes a drain which being connected correspondingly to one of the plurality of drain select input lines, e.g., DSEL<sub>0</sub>, DSEL<sub>1</sub>, etc., for receiving one of the mutually exclusive drain select signals therefrom. Each of the transistors further includes a gate which being connected correspondingly to one of the plurality of gate select input lines GSEL<sub>0</sub>, GSEL<sub>1</sub>, etc., for receiving one of the mutually exclusive gate select signals therefrom. Each of the plurality of transistors further includes a source, which is connected to an output signal line for providing the activating signal to one of the word lines WL which being contingent upon the mutually exclusive drain select signals DSEL<sub>0</sub>, DSEL<sub>1</sub>, etc. And, the mutually exclusive gate select signals GSEL<sub>0</sub>, GSEL<sub>1</sub>, etc. In a preferred embodiment, each of the transistors is an enhanced mode transistor, and in another preferred embodiment, each of the transistors is a depletion mode transistor.

Furthermore, according to FIGS. 3a to 4b, and FIGS. 12(a) to 13 a method for configuring a semiconductor

6,108,229

19

memory device for operation with a plurality of cell-refreshing sense-amplifiers (SAs) is also disclosed. The method includes the steps of (a) arranging a plurality of first-direction first-level bit lines in a parallel manner along a first direction; (b) arranging a plurality of word lines for intersecting with the first-direction first-level bit lines; (c) coupling a memory cell between each of the first-direction first level bit lines and one of the word lines for storing data therein; (d) arranging a plurality of different-direction first level bit lines along a plurality of different directions with at least one of the different directions being different from the first direction; (e) connecting each of the first direction first level bit lines to one of the cell-refreshing sense amplifiers (SAs) directly or via the different-direction first level bit lines; (f) connecting each of the word lines WL to a decoder 302 for receiving an activating signal therefrom for activating one of the word lines WL; (g) forming the decoder with a plurality of transistors each includes a drain, a gate and a source therein; (h) connecting a drain select line to each of the drain of each of the transistors and connecting a gate select line to each of the gate of each of the transistors; (i) applying each of the drain select lines to receive one of a plurality of mutually exclusive drain select signals and each of the gate select lines to receive one of a plurality of mutually exclusive gate select signals; and (j) applying each of the plurality of transistors to generate an output signal from each of the source which being contingent upon the mutually exclusive drain select signals and the mutually exclusive gate select signals for providing the activating signal to each of the word lines.

According to FIG. 13, this invention further discloses a dynamic random access memory (DRAM) cell which is coupled to a word-line and a bit-line. The DRAM memory cell includes a select transistor 1402 includes a drain connected to the bit line BL and a gate connected to the word line WL. The cell further includes a storage transistor 1404 includes a drain connected to the source of the select transistor 1402 and a gate connected to a power supply voltage Vcc whereby the storage transistor 1404 is implemented as a capacitor for storing a binary bit therein. In summary, the present invention further discloses a memory cell coupled to a word-line and a bit-line. The memory cell includes a storage transistor connected to the word line and bit line via a select means provided for selectively activating the memory cell. And, the storage transistor further includes a gate, which is biased to a power supply voltage to function, as a capacitor for storing a binary bit therein.

FIGS. 14(a-f) and FIGS. 15(a-c) illustrates a procedure to manufacture high density memory using a manufacture technology very similar to standard logic technology. The first step is to define active area 1502, and grow isolation field oxide 1504 to separate those active area as show in the cross section diagram in FIG. 14(a) and the top view in FIG. 15(a). This step is identical to any standard IC technology. The next step is to apply a mask 1506 to define the location of trench capacitors as illustrated by FIG. 14(b). Selective plasma etching is used to dig a trench 1510 at the opening defined by the field oxide 1504 and the trench mask 1506 as illustrated in the cross-section diagram in FIG. 14(c) and the top view in FIG. 15(b). This is a self-aligned process because three edges of the trench 1510 are defined by field oxide. The trench mask 1506 only needs to define one edge of the trench. After the above processing steps, all the following processing procedures are conventional manufacture processes of standard logic technology. First, a layer of thin insulator 1511 is grown at the surface of the active area 1502, including the surfaces of the trench 1510 as shown in

20

FIG. 14(d). The next step is to deposit poly silicon 1512 to fill the trench 1510 and cover the whole silicon as illustrated in FIG. 14(e). A poly mask 1520 is then used for poly silicon etching process to define transistor gates 1522 and the electrode 1524 of the trench capacitor as illustrated in FIG. 14(f). FIG. 15(c) shows the top view and FIG. 15(g) shows the cross-sectional view of the resulting memory cell structure. The trench capacitors 1510 are filled with poly silicon. One electrode 1602 of all those trench capacitors 1510 are connected together through poly silicon to power supply voltage Vcc. The other electrodes of the trench capacitors are connected to the sources of select transistors 1604. The poly silicon word lines 1606 define the gates of the select transistors, and the drains of the select transistors are connected to metal bit lines through diffusion contacts 1608.

As apparent from the foregoing, following advantages are obtained according to this invention.

(1) All the procedures used to build the DRAM cell are existing procedures of standard logic technology, except one masking step and one plasma-etching step. Comparing with current art embedded memory technologies, the present invention simplifies the manufacture technology by more than 30%.

(2) The procedure to define the dimension of trench capacitor is a self-aligned procedure; three edges of the trench capacitor are defined by field oxide; only one edge is defined by mask. This self-aligned procedure allows us to minimize the area of the memory cell.

Another procedure has also been developed to build self-aligned trench capacitors using logic technology. The first step is to build CMOS transistors following standard logic technology as illustrated in the cross-section diagram in FIG. 16(a). At this time, the MOS transistor has been fully processed. The poly silicon gate 1702 is already covered by oxide for protection. A trench mask 1706 is then deposited. This trench mask 1706 is used to protect area where we do not want to dig trench capacitor, it is not needed to define the dimension of the trench capacitor because all four edges of the area are already defined. Three edges are defined by the field oxide 1710 in the same way as the previous procedure, and the forth edge is define by the oxide 1704 on the transistor gate. This is therefore a complete self-aligned procedure. The following selective plasma etching procedure is therefore able to utilize optimum area for the trench capacitor as illustrated in FIG. 16(b). Thin insulation layer is grown on the surfaces of the trench 1712 before the whole area is covered by second layer poly silicon 1714 as shown in FIG. 16(c). Photo resist 1716 that is defined by the same mask as the one used in FIG. 16(a) defines the dimension of the second layer poly silicon 1716 (the polarity of the photo resist used in FIG. 16(a) is opposite to that used in FIG. 16(c). The second layer poly silicon 1716 is then etched to form the electrodes 1720 of those trench capacitors 1722. FIG. 17 shows the top view of the DRAM cells manufactured by the above procedures. The word lines 1802 are defined by the first layer poly silicon. Second layer poly silicon are used to fill the trench capacitors 1722, and to connect one electrode 1720 of all those trench capacitors to Vcc.

The above procedure is more complex than the procedure illustrated in FIGS. 14(a-g). It has the advantage that the trench capacitors are fully self-aligned for all 4 edges of their opening. Utilization of the silicon area is therefore fully optimized. While specific embodiments of the invention have been illustrated and described herein, it is realized that other modification and changes will occur to those skilled in the art. For example, the insulation-layer in the trench



6,108,229

21

capacitors maybe grown in a different processing step instead of during the process of forming the gate oxide. The exact sequence of the processing steps also can be varied to achieve similar simplification.

The top erode (1602) of the trench capacitor (1510) of the memory cells shown in FIG. (14) must be connected to a voltage at least one threshold voltage ( $V_t$ ) higher than the voltage of the bottom electrode to make the area under the insulator layer (1511) conductive. Similarly, the top electrode (1702) of the trench capacitor of the memory cells shown in FIG. (16) also must be connected to a voltage at least one  $V_t$  higher than the voltage of bottom electrode. Typically, those top electrodes (1602, 1702) are connected to power supply voltage  $V_{cc}$ . This constraint can be removed if a diffusion layer (1805) is deposited around the trench capacitor (1802) as illustrated by the cross-section diagram in FIG. 18(a). This diffusion layer (1805), the drain of the word line transistor (1606), and the top electrode (1602) are all doped with the same type of doping. Therefore, the bottom electrode of the trench capacitor (1801) is always conductive, which removes the constraint on the electrode voltages. The cross-section diagram in FIG. 18(b) illustrates another variation in device structure. In this structure, a transistor (1811) instead of field oxide separates two nearby trench capacitors (1821, 1823). The gate (1813) of this isolation transistor (1811) is connected to ground voltage  $V_{ss}$  to separate nearby trench capacitors (1821, 1823). Transistors (1811, 1815) therefore define two edges of the areas of the trench capacitors (1821, 1823) instead of field oxide, which usually helps to reduce the size of memory cells.

In the above examples, the geometry of memory cell structures is drawn in 90-degree angles for simplicity. In reality, memory cells are often drawn in multiple angles as illustrated by the top view memory cell structures in FIG. 19. The trench capacitors (1901) are placed in 45 degree to the contacts (1903). The word line (1907) and the diffusion area (1905) are also placed in 45-degree angles. Since the area of the trench capacitors (1901) are defined by field oxide and transistor edges, its shape is therefore not necessary rectangular as shown by the example in FIG. 19.

The word line transistor (1402) in the memory cell of the present invention has the same properties and it is manufactured in the same time as the transistors used for peripheral circuits and logic circuits. The word line transistors of prior art DRAM technologies are always different from logic transistors. In order to tolerate higher word line voltage introduced by the word line boosting circuits, the gate oxide thickness ( $T_{ox}$ ) of a prior art word line transistor is thicker than that of a logic transistor. In order to reduce leakage current, the threshold voltage ( $V_t$ ) of a prior art word-line-transistor is higher. Table 1 lists transistor properties for a typical 0.35  $\mu m$  DRAM technology. The word line transistor and the logic transistor in this example is manufactured by the same procedures except that one masking step is added to increase  $V_t$  of the word line transistor. The word line transistor has higher  $V_t$  (1.1 volts for the example in Table 1) so that it can be drawn to a smaller minimum channel length ( $L_{min}$ ), which is 0.35  $\mu m$  in this case, without leakage problems. The logic transistor has lower  $V_t$  (0.7 volts for this example), but its  $L_{min}$  is larger. On the other word, the logic transistors of a typical DRAM technology is equivalent to the logic transistors of 0.5  $\mu m$  technology instead of 0.35  $\mu m$  technology. On the other word, the performance of logic transistors of DRAM technology is one generation behind the transistors of typical logic technology.

One method to have both high performance logic transistors and low leakage DRAM transistors on the same chip

22

is to make different kinds of transistors using complex manufacture procedures. Table 2 shows the transistor properties for one example of such complex embedded memory technology. This technology has word line transistor with high  $V_t$  and thick oxide, high voltage transistors with thick oxide and long channel length, and logic transistors with low  $V_t$  and thin oxide. The manufacture procedures for such technology are very complex. The manufacture cost is very high.

TABLE 1

Transistor properties for word line transistors and logic transistors of prior art DRAM technology.			
	$T_{ox}$	$V_t$ (volts)	$L_{min}$ (micrometers)
Word line transistor	100	1.1	0.35
Logic transistor	100	0.7	0.5

TABLE 2

Transistor properties for word line transistors and logic transistors of prior art embedded DRAM technology.			
	$T_{ox}$	$V_t$ (volts)	$L_{min}$ (micrometers)
Word line transistor	100	1.1	0.35
High Voltage transistor	100	0.7	0.5
Logic transistor	70	0.7	0.35

TABLE 1

Transistor properties for word line transistors and logic transistors of prior art DRAM technology.			
	$T_{ox}$	$V_t$ (volts)	$L_{min}$ (micrometers)
Word line transistor	100	0.7 (1.1)	0.35
Logic transistor	100	0.7	0.35

A DRAM (dynamic random access memory) cell array supported on a substrate is therefore disclosed in this invention. The DRAM cell array includes a plurality of memory cells each having a select-transistor wherein each of the select-transistor having a select-transistor-gate. The DRAM cell array further includes a peripheral logic-circuit having logic-transistors wherein each of the logic-transistors having a logic-transistor-gate. The select-transistor-gate and the logic-circuit-gate have substantially a same thickness. And, the select-transistor for each of the memory cells having a select-transistor threshold voltage and each of the logic-transistors of the peripheral logic-circuit having a logic-transistor threshold voltage wherein the select-transistor threshold voltage is substantially the same as the logic-transistor threshold voltage. In a preferred embodiment, each of the memory cells further having a trench capacitor. In another preferred embodiment, the DRAM cell array further includes an active area isolated and defined by edges of a field oxide layer disposed on the substrate wherein each of the trench capacitors disposed in the active area and in self-alignment with the edges of the field oxide layer. In another preferred embodiment, the DRAM cell array further includes an active area isolated and defined by edges of a

6,108,229

23

field oxide layer disposed on the substrate. Each of the trench capacitors is disposed in the active area and in self-alignment with the edges of the field oxide layer and edges of the select-transistor gate. In another preferred embodiment, the DRAM cell array further includes an error code checking (ECC) and correction means connected to the memory cell array for checking and correcting substantially all memory read errors within a threshold error-detection-and-correction time.

According to above description, this invention discloses a method for manufacturing a DRAM (dynamic random access memory) cell array each having a select-transistor and peripheral logic circuit having logic-transistors supported on a substrate. The method includes the steps of (a) applying a gate-formation process for simultaneously forming a select-transistor-gate for the select-transistor and a logic-circuit-gate for each of the logic-transistors for the peripheral logic-circuit wherein the select-transistor-gate and the logic-circuit-gate having substantially a same thickness; and (b) applying substantially same implant processes in forming the select-transistor and the logic-transistors wherein the select-transistor and the logic transistors having substantially a same threshold voltage. In a preferred embodiment, the method further includes a step of (c) applying a capacitive-transistor trench mask for etching a plurality of trench capacitors for the memory cell array. In a preferred embodiment, the step of applying a capacitive-transistor trench mask is a step of applying a capacitive-transistor trench mask in an active area isolated by a field oxide. The capacitive-transistor trench mask cooperates with the filed oxide for etching the trench in self-alignment in the active area with etching edges defined by the field oxide. In another preferred embodiment, the step of applying a capacitive-transistor trench mask in corporation with the field oxide is a step of applying a capacitive-transistor trench mask in an active area isolated by the field oxide as an enclosed area. The capacitive-transistor trench mask is employed to define a single edge of the trench capacitor while remaining edges of the trench capacitor are in self-alignment with the field oxide wherein the etching edges for the remaining edges are inherently defined in the active area by the filed oxide. In another preferred embodiment, the step of applying a capacitive-transistor trench mask in corporation with the field oxide is a step of applying a capacitive-transistor trench mask in an active area isolated as an enclosed area by the filed oxide and a gate in the active area. The capacitive-transistor trench mask is employed to define a single edge of the trench capacitor while remaining edges of the trench capacitor are in self-alignment with the field oxide and the gate. The etching edges for the remaining edges are inherently defined in the active area by the field oxide and the gate. In a preferred embodiment, the method further includes steps of: (d) removing the capacitive-transistor trench mask after etching the trench capacitor followed by filling the capacitor trench with a layer of polycrystalline silicon overlaying the active area; and (e) applying the capacitive-transistor trench mask again in opposite polarity relative to the step described above to etch the polycrystalline layer to define a contact opening to the trench capacitor.

According to above drawings and descriptions, this invention also discloses a method for manufacturing a DRAM (dynamic random access memory) cell array on a substrate. The method includes the steps of (a) forming logic transistors on the substrate having polysilicon gates covered by an insulation protective layer wherein the insulation protective layer disposed next to a field oxide layer defining open areas

24

therein-between; and (b) forming trench capacitors for the memory cells by etching the open areas with edges of the trenches defined by the insulation protective layer and the field oxide layer. In a preferred embodiment, the step of forming logic transistors on the substrate having polysilicon gates comprising a step of forming word-line (WL) select transistors each having a WL-transistor gate padded with a WL-select gate-oxide layer having a thickness substantially the same as a gate oxide layer padded under the polysilicon gates of the logic transistors. In another preferred embodiment, the method further includes a step of (c) connecting an error code checking (ECC) and correction means to the memory cell array for checking and correcting substantially all memory read errors within a threshold error-detection-and-correction time. In another preferred embodiment, the method further includes a step of (e) forming a diffusion layer surrounding the trenches having a same conductivity type as a drain of the logic transistors. In another preferred embodiment, the method further includes a step of (f) forming logic transistors on the substrate having polysilicon gates covered by an insulation protective layer; (f) connecting the gate of a plurality of the logic transistors to a ground voltage thus defining a plurality of isolation transistors each separating two adjacent logic transistors wherein the insulation protective layer of the isolation transistors and the adjacent logic transistors defining open areas therein-between; and (g) forming trench capacitors for the memory cells by etching the open areas with edges of the trenches defined by the insulation protective layer of the isolation transistors and the adjacent logic transistors.

An embedded technology of the present invention uses high performance transistor to support both logic circuits and memory circuits. The circuit performance is high, and the manufacture procedures are simple. However, the leakage current caused by the word line transistor is higher than that of prior art word line transistor. Since the thin gate device can not tolerate high voltage operation, we can not use word line boost method to increase storage charge. It is therefore necessary to provide novel design methods to improve the tolerance in leakage current and storage charge loss. U.S. Pat. No. 5,748,547 disclosed methods that can improve signal-to-noise ratio of DRAM array without increasing device area. Using the method, memory devices can be functional without using boosted word line voltages. The same patent disclosed novel self-refresh mechanism that is invisible to external users while using much less power. Using the self-refresh mechanism to increase refresh frequency internally, we can tolerate higher memory leakage current without violating existing memory specifications. Another important method is to use the error-correction-code (ECC) protection to improve the tolerance in non-ideal memory properties.

FIG. 20(a) shows a typical distribution for the refresh time required by the memory cells in a large memory device. For a prior art memory device, the refresh time of the worst bit, i.e., ( $T_{min}$ ), determines its refresh time, among millions of memory cells in the memory device. This worst bit refresh time ( $T_{min}$ ) is typically many orders of magnitudes shorter than the average refresh time ( $T_{av}$ ), because the worst bit is always caused by defective structures in the memory cell. FIG. 20(b) shows the simplified block diagram of a memory device equipped with ECC protection circuits. During a memory write operation, the input data is processed by a ECC parity tree (2005) to calculate ECC parity data. The input data is stored into a normal data memory array (2001) while the ECC parity data is stored into a parity data array (2003). During a read operation, stored data as well as ECC

6,108,229

25

parity data are read from the memory arrays (2001, 2003) and sent to the ECC parity tree (2005). In case there are corruption data, an ECC correction logic (2007) will find out the problem and correct the error so that the output data will be correct. The ECC correction mechanism is known to the art, but it has not been used on low-cost DRAM because it will require more area. The present invention use ECC protection as a method to improve the tolerance in memory cell leakage current. When a memory device is equipped with an ECC circuit, it will correct most single-bit errors. As a result, the refresh time of the memory device is no longer dependent on the worst bit in the memory. Instead, the device will be function until the errors are more than what the ECC mechanism can correct. The refresh time (T<sub>ecc</sub>) is therefore higher than T<sub>min</sub> as shown in FIG. 20(a).

Base on the above novel design methods, practical memory devices using high performance logic transistor in DRAM memory cells have been manufactured successfully.

Although the present invention has been described in terms of the presently preferred embodiment, it is to be understood that such disclosure is not to be interpreted as limiting. Various alternations and modifications will no doubt become apparent to those skilled in the art after reading the above disclosure. Accordingly, it is intended that the appended claims be interpreted as covering all alternations and modifications as fall within the true spirit and scope of the invention.

I claim:

1. A DRAM (dynamic random access memory) cell array supported on a substrate comprising:

a plurality of memory cells each having a select-transistor wherein each of said select-transistor having a select-transistor-gate;

26

a peripheral logic-circuit having logic-transistors wherein each of said logic-transistors having a logic-transistor-gate;

said select-transistor-gate and said logic-circuit-gate having substantially a same thickness;

said select-transistor for each of said memory cells having a select-transistor threshold voltage and each of said logic-transistors of said peripheral logic-circuit having a logic-transistor threshold voltage wherein said select-transistor threshold voltage is substantially the same as said logic-transistor threshold voltage.

2. The DRAM cell array of claim 1 wherein:

each of said memory cells further having a trench capacitor.

3. The DRAM cell array of claim 2 further comprising: an active area isolated and defined by edges of a field oxide layer disposed on said substrate wherein each of said trench capacitors disposed in said active area and in self-alignment with said edges of said field oxide layer.

4. The DRAM cell array of claim 2 further comprising: an active area isolated and defined by edges of a field oxide layer disposed on said substrate wherein each of said trench capacitors disposed in said active area and in self-alignment with said edges of said field oxide layer and edges of said select-transistor gate.

5. The memory cell array of claim 1 further comprising: an error code checking (ECC) and correction means connected to said memory cell array for checking and correcting substantially all memory read errors within a threshold error-detection-and-correction time.

\* \* \* \* \*

# EXHIBIT B



(12) **United States Patent**  
**Shau**

(10) **Patent No.:** **US 6,687,148 B2**  
(45) **Date of Patent:** **Feb. 3, 2004**

(54) **HIGH PERFORMANCE EMBEDDED  
SEMICONDUCTOR MEMORY DEVICES  
WITH MULTIPLE DIMENSION FIRST-  
LEVEL BIT-LINES**

(51) **Int. Cl.<sup>7</sup>** ..... **G11C 11/24**  
(52) **U.S. Cl.** ..... **365/63; 365/149; 257/301**  
(58) **Field of Search** ..... **365/63, 51, 72,**  
**365/149, 150; 257/296, 297, 301**

(75) **Inventor:** **Jeng-Jye Shau, Palo Alto, CA (US)**

(56) **References Cited**

(73) **Assignee:** **UniRam Technology, Inc., Santa Clara,  
CA (US)**

**U.S. PATENT DOCUMENTS**

4,926,223 A \* 5/1990 Bergemont ..... 365/149  
5,946,230 A \* 8/1999 Shimizu et al. .... 365/149  
6,339,240 B1 \* 1/2002 Kim ..... 365/149

(\*) **Notice:** Subject to any disclaimer, the term of this  
patent is extended or adjusted under 35  
U.S.C. 154(b) by 0 days.

\* cited by examiner

*Primary Examiner*—Tan T. Nguyen

(74) *Attorney, Agent, or Firm*—Bo-In Lin

(21) **Appl. No.:** **10/269,571**

(57) **ABSTRACT**

(22) **Filed:** **Oct. 10, 2002**

(65) **Prior Publication Data**

US 2003/0043657 A1 Mar. 6, 2003

**Related U.S. Application Data**

(63) Continuation of application No. 09/860,215, filed on May  
18, 2001, now Pat. No. 6,504,745, which is a continuation-  
in-part of application No. 08/805,290, filed on Feb. 25,  
1997, now Pat. No. 5,825,704, and a continuation-in-part of  
application No. 08/653,620, filed on May 24, 1996, now Pat.  
No. 5,748,547.

A dynamic random access memory solves long-existing  
tight pitch layout problems using a multiple-dimensional bit  
line structure. Improvement in decoder design further  
reduces total area of this memory. A novel memory access  
procedure provides the capability to make internal memory  
refresh completely invisible to external users. By use of such  
memory architecture, higher performance DRAM can be  
realized without degrading memory density. The require-  
ments for system support are also simplified significantly.

**31 Claims, 30 Drawing Sheets**

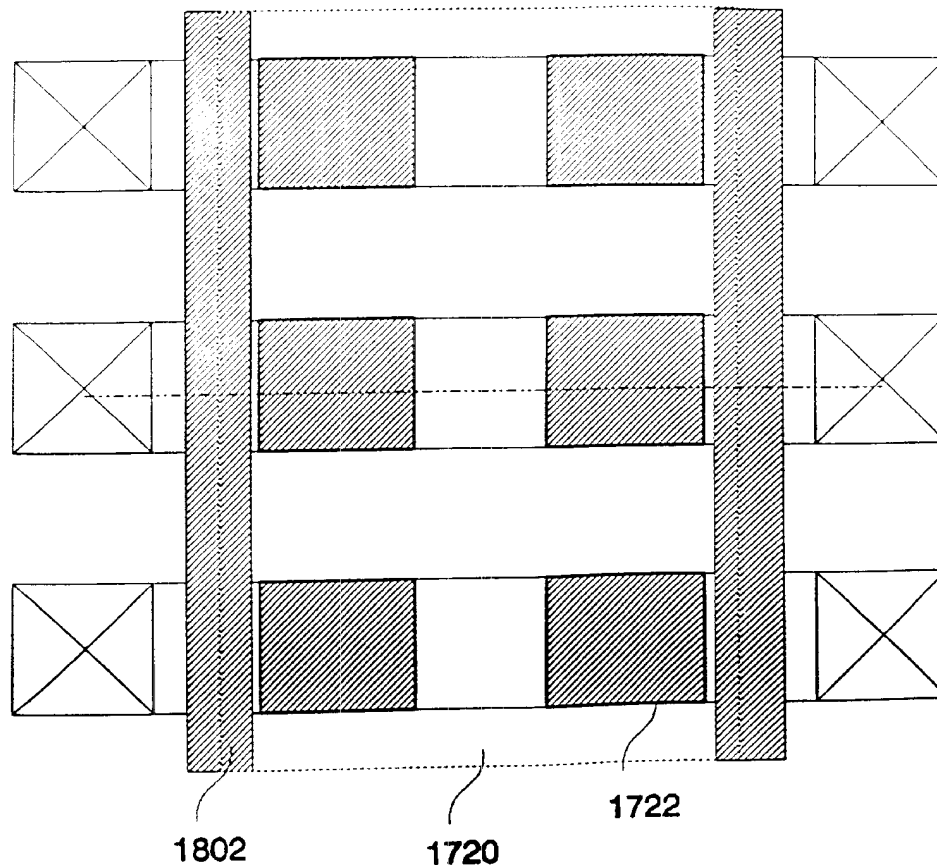


FIG.1 (Prior Art)

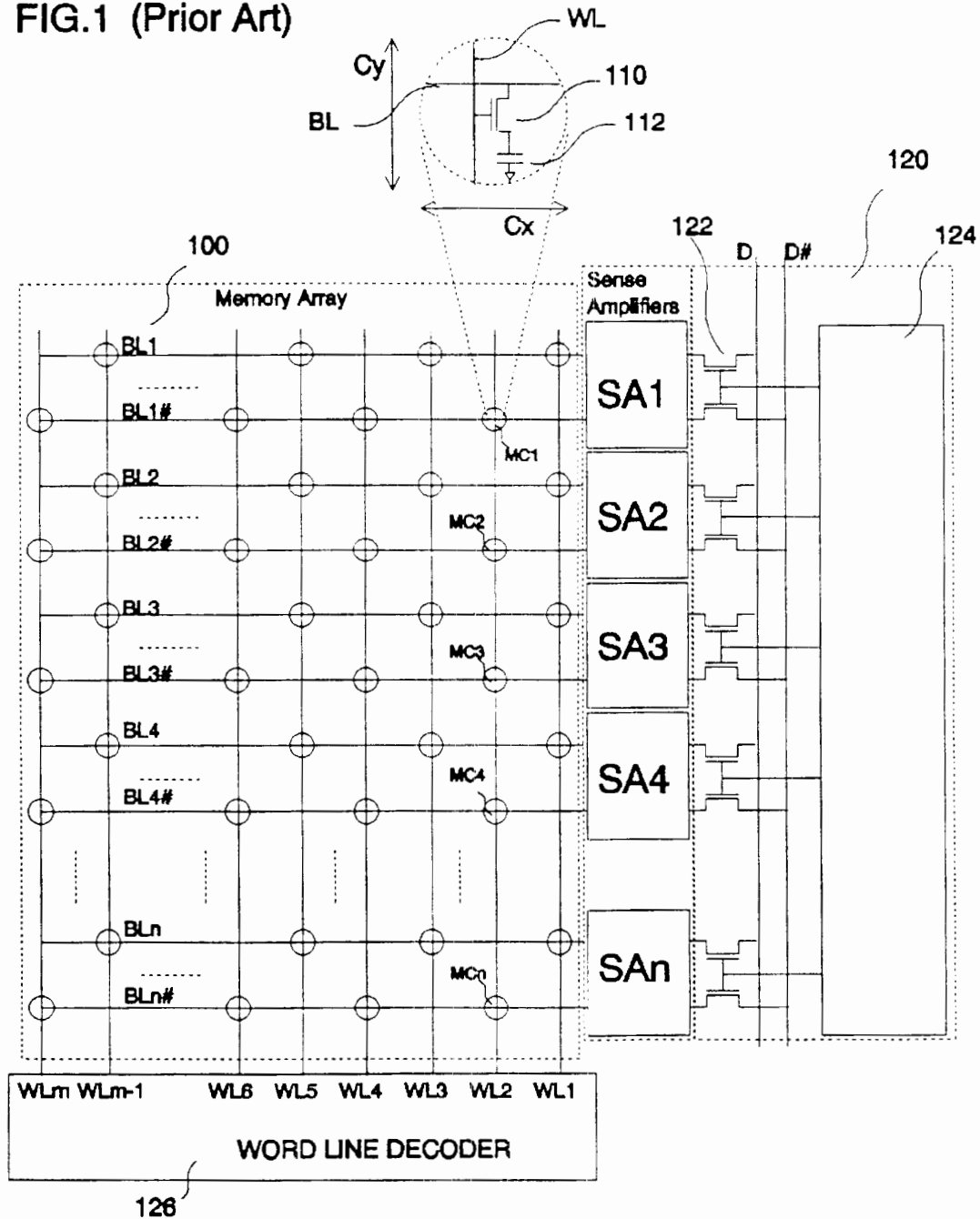


FIG. 2 (prior art multi-bank memory)

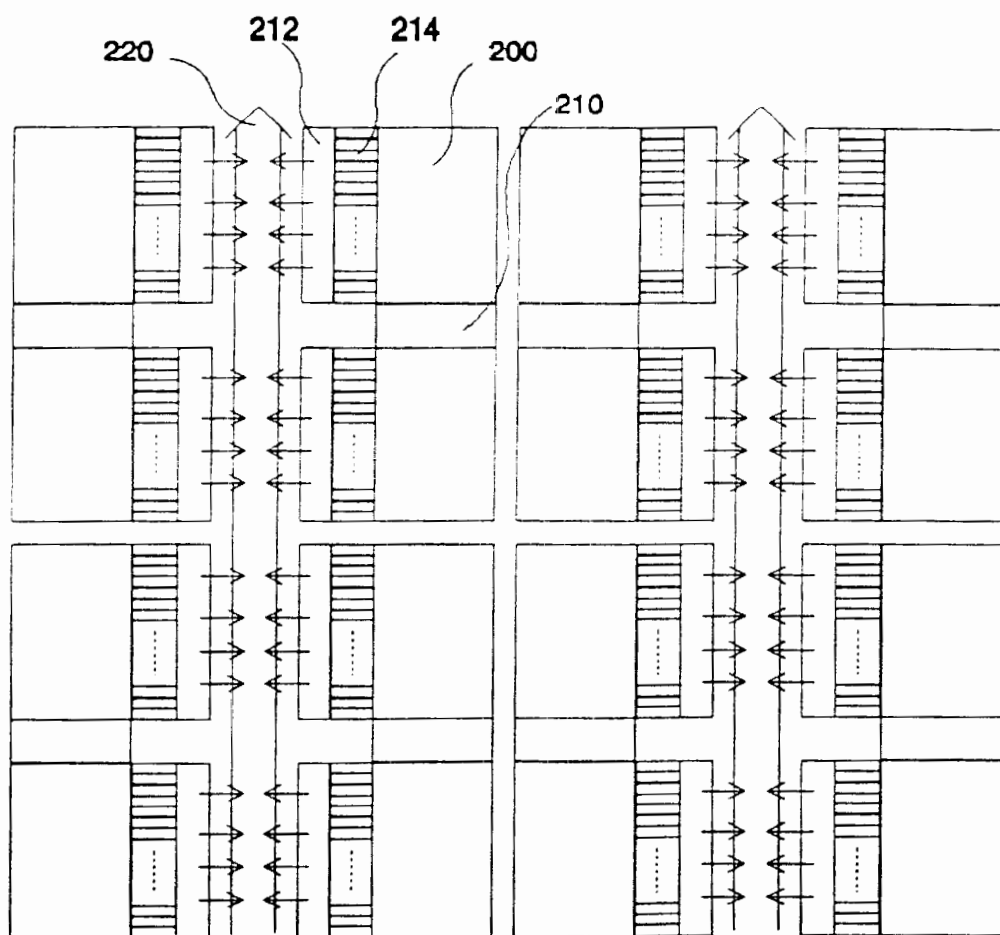


FIG. 3a

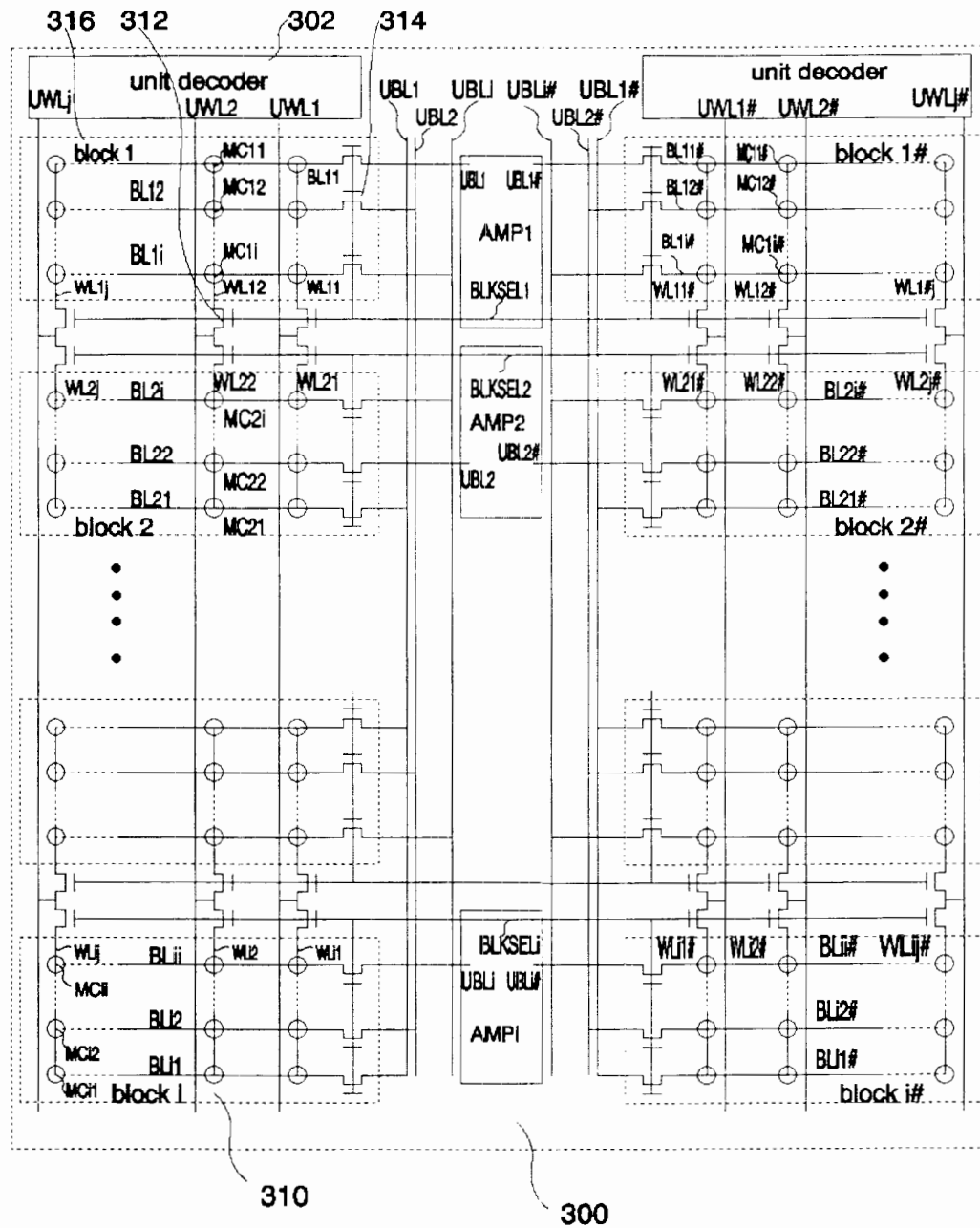


FIG. 3b

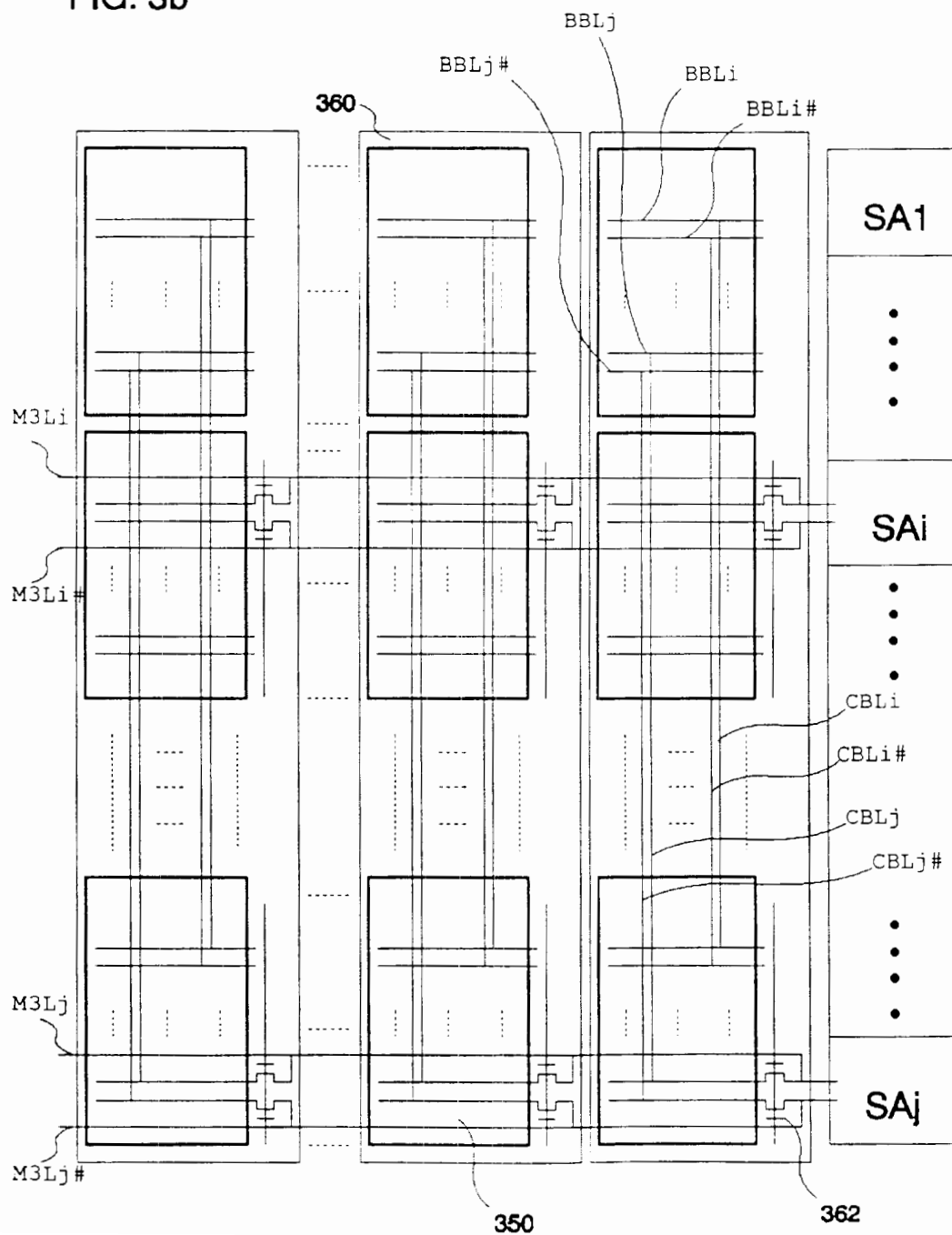


FIG. 4a (prior art)

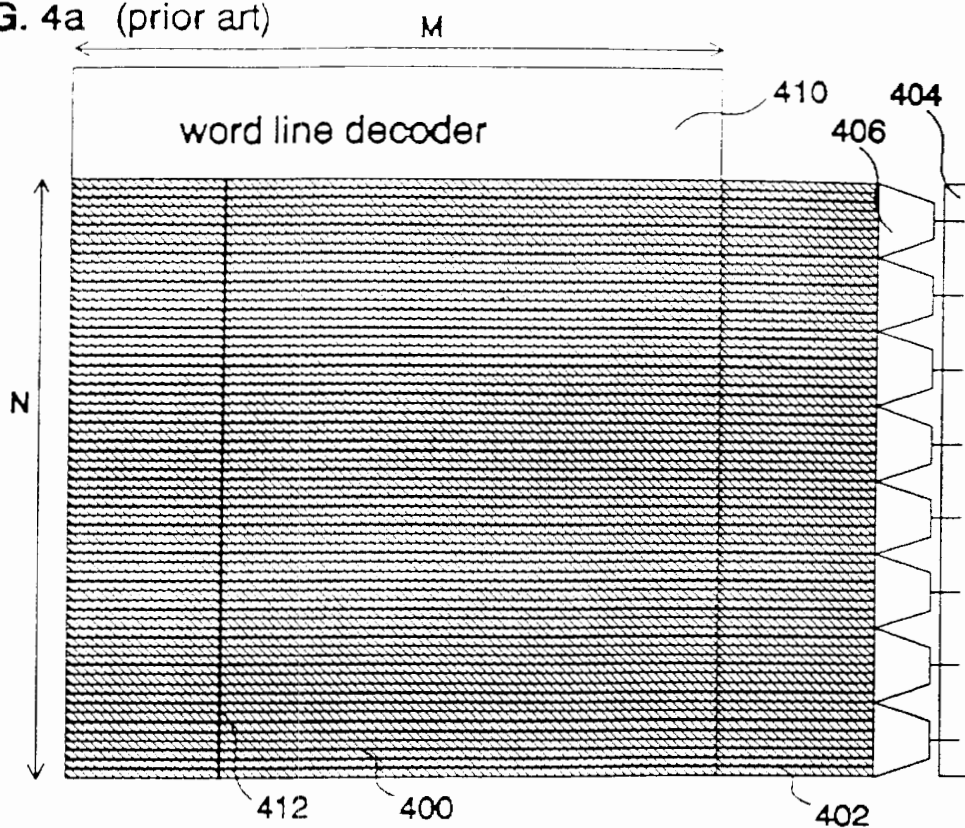


FIG. 4b

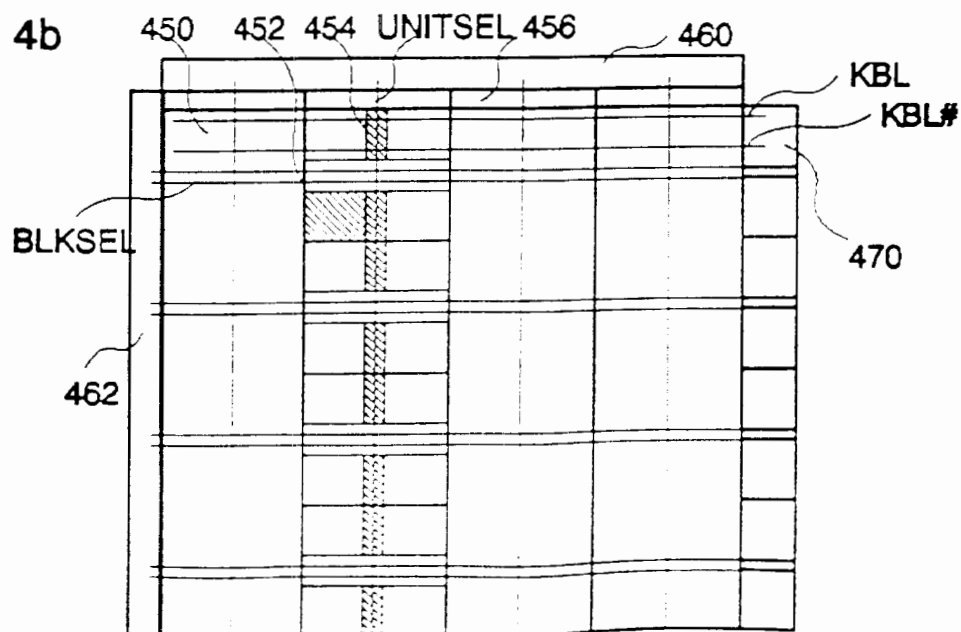


FIG. 5

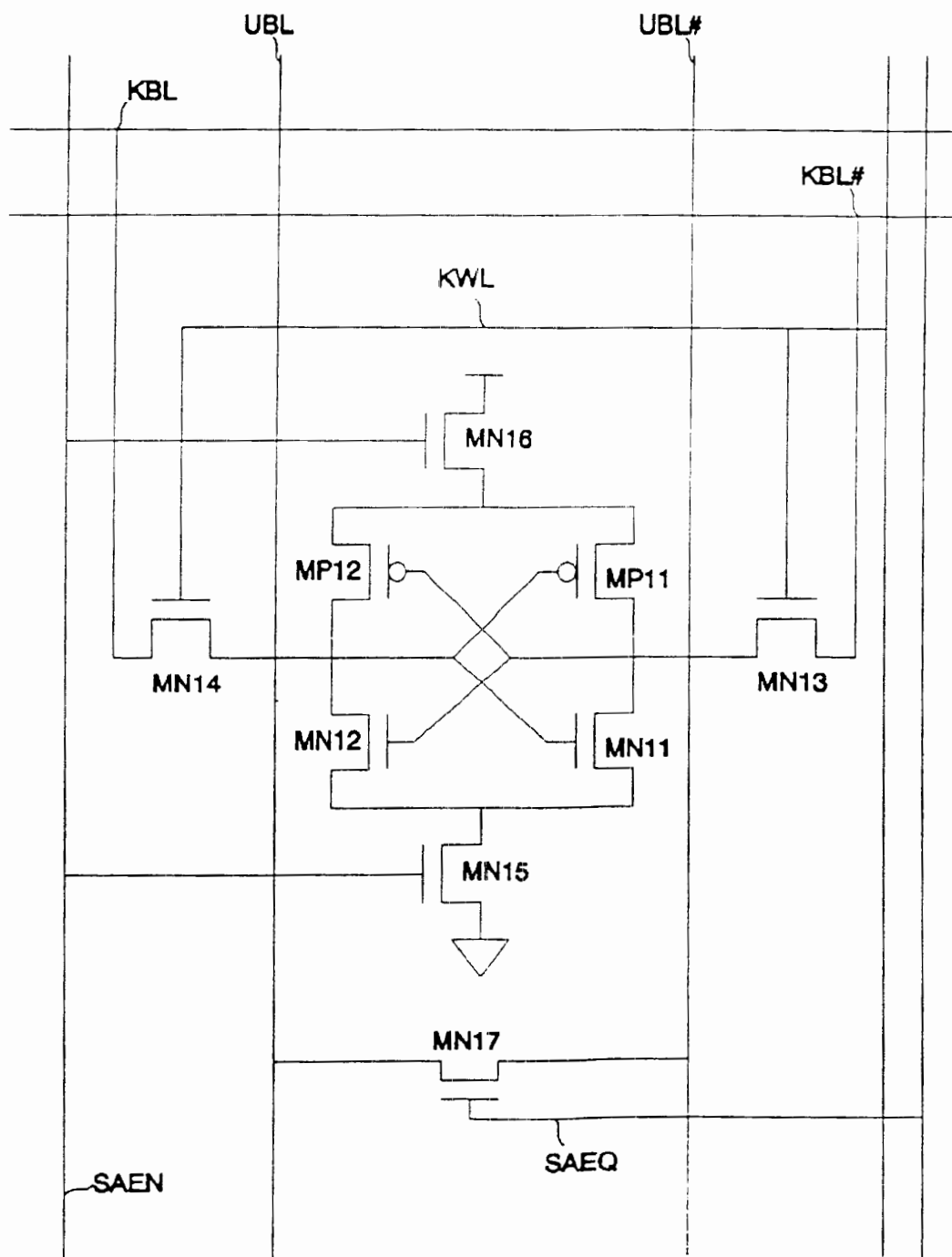




FIG. 6

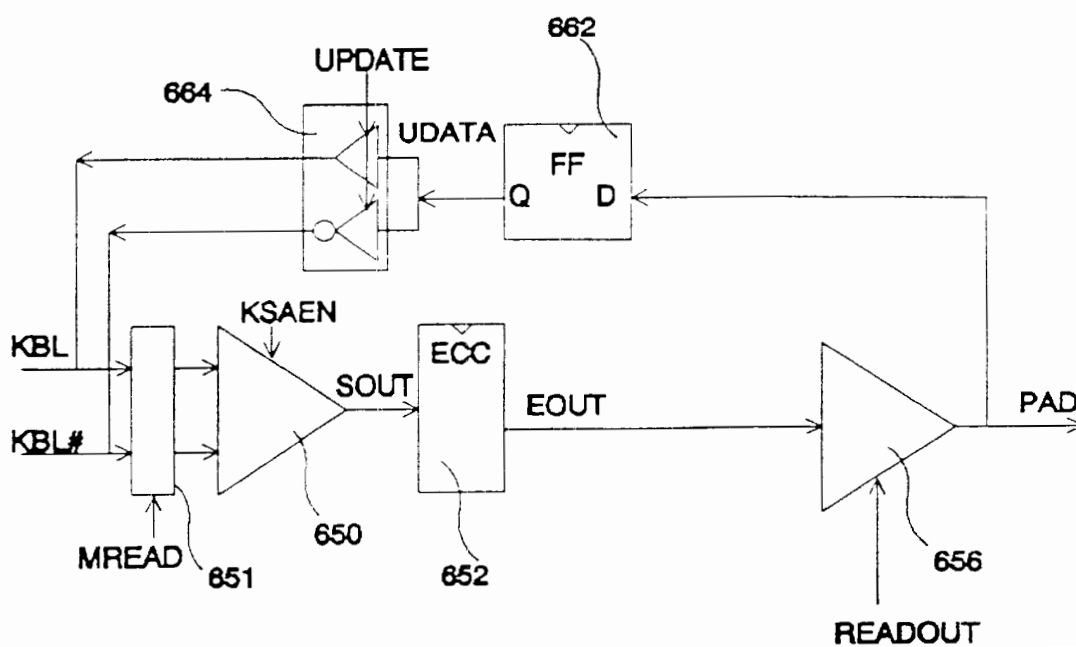


FIG. 7a (read cycle)

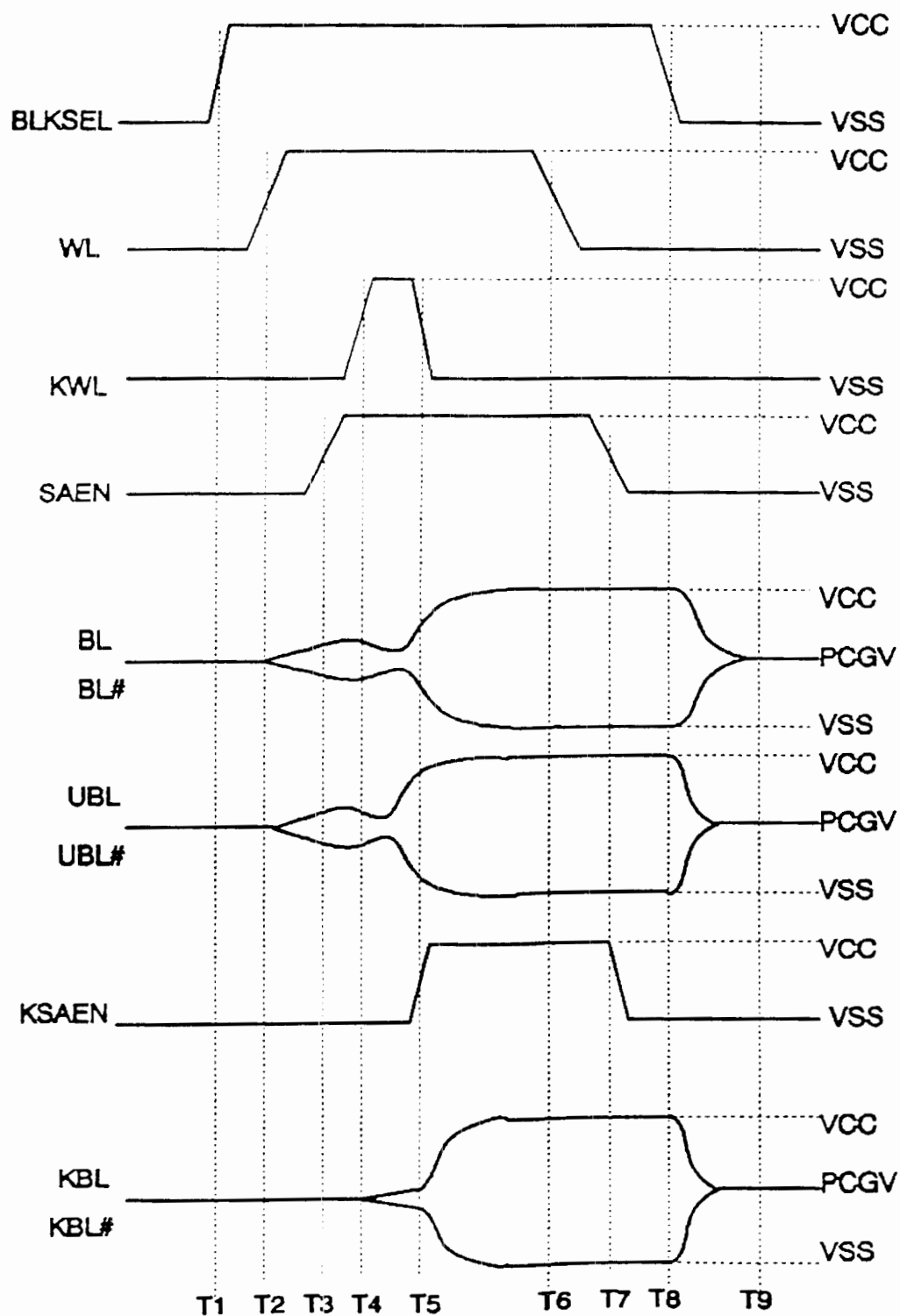


FIG. 7b (refresh cycle)

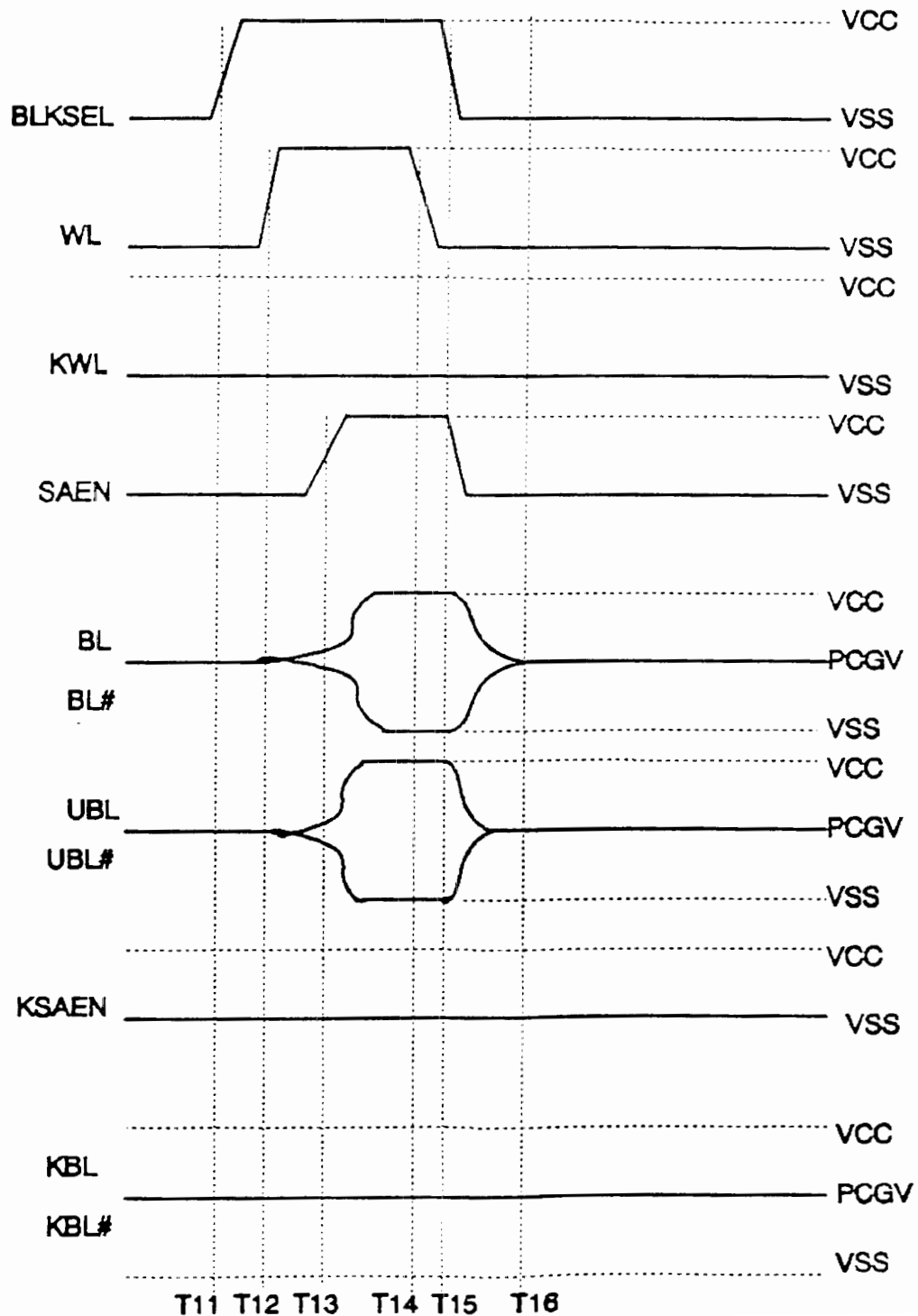


FIG. 7c (update cycle)

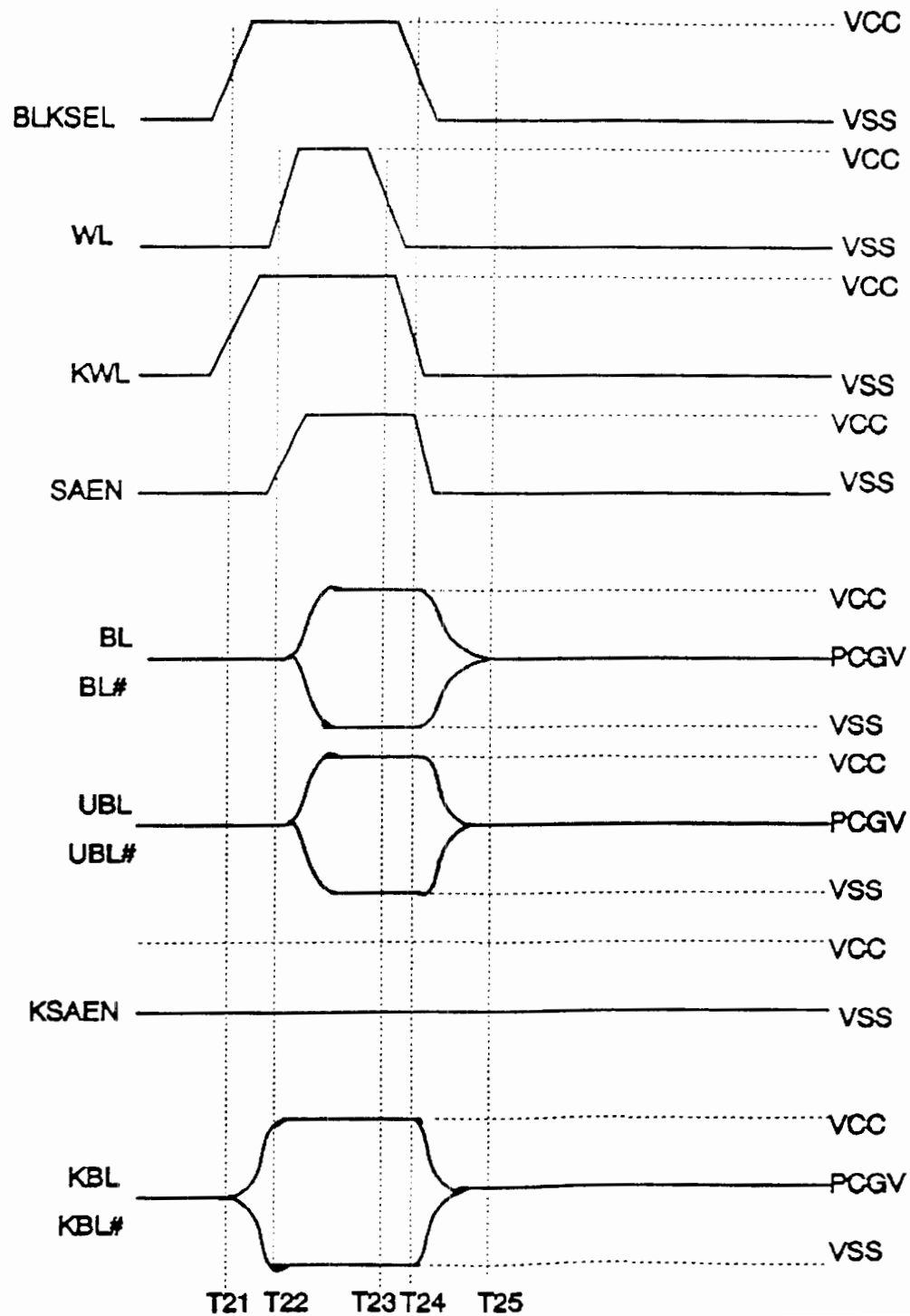


FIG. 8

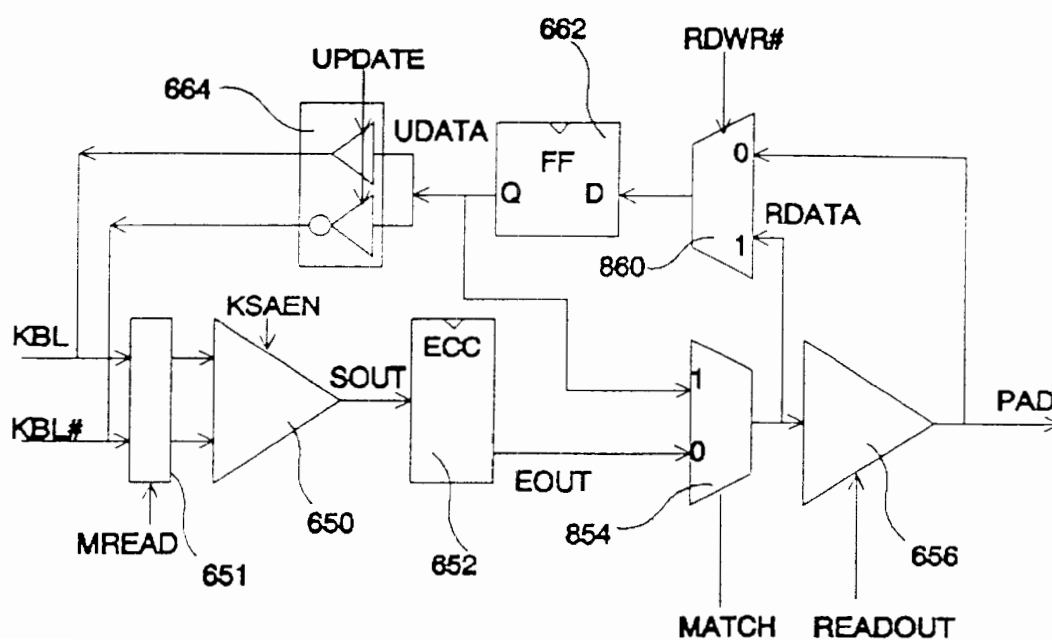


FIG. 9

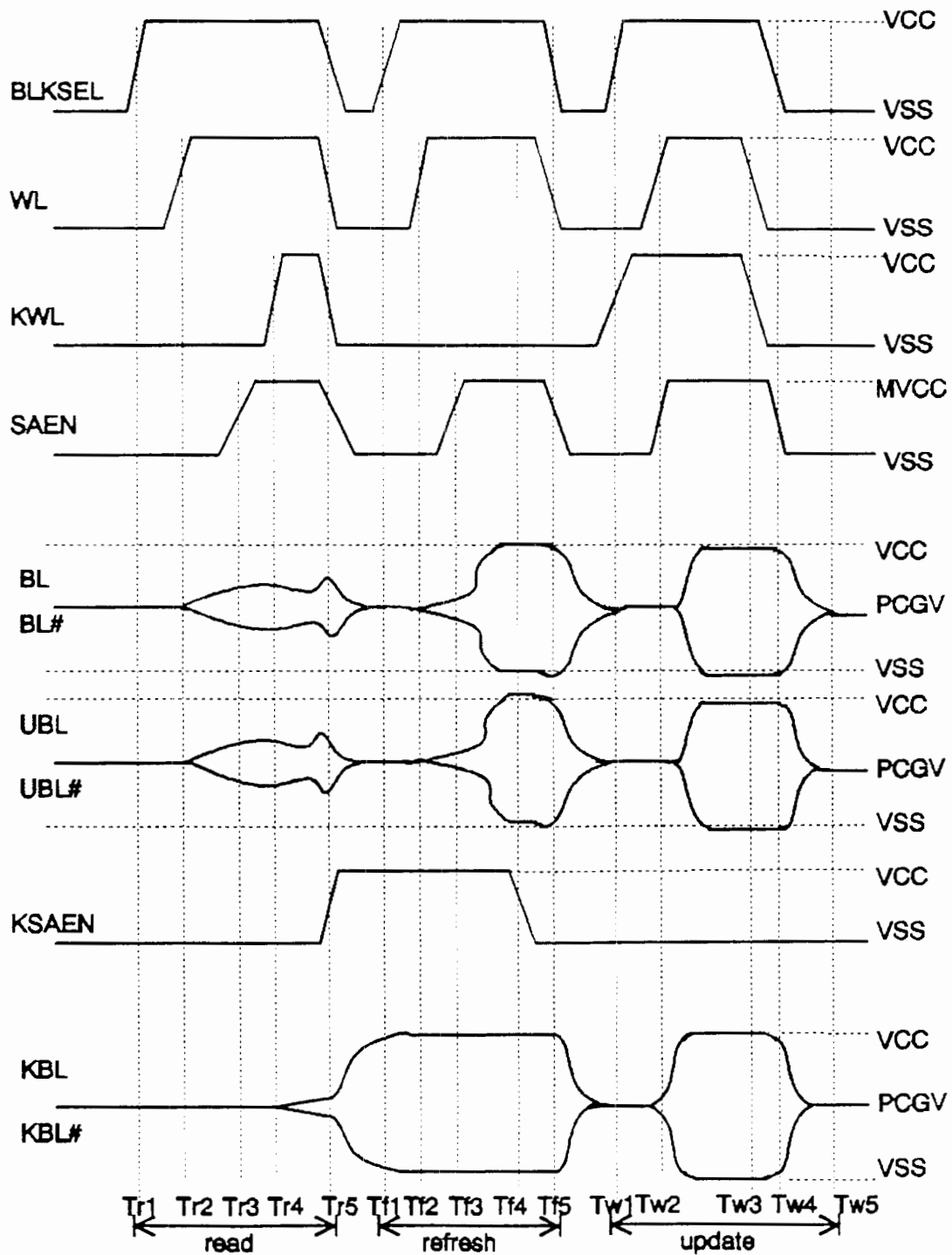




FIG. 10

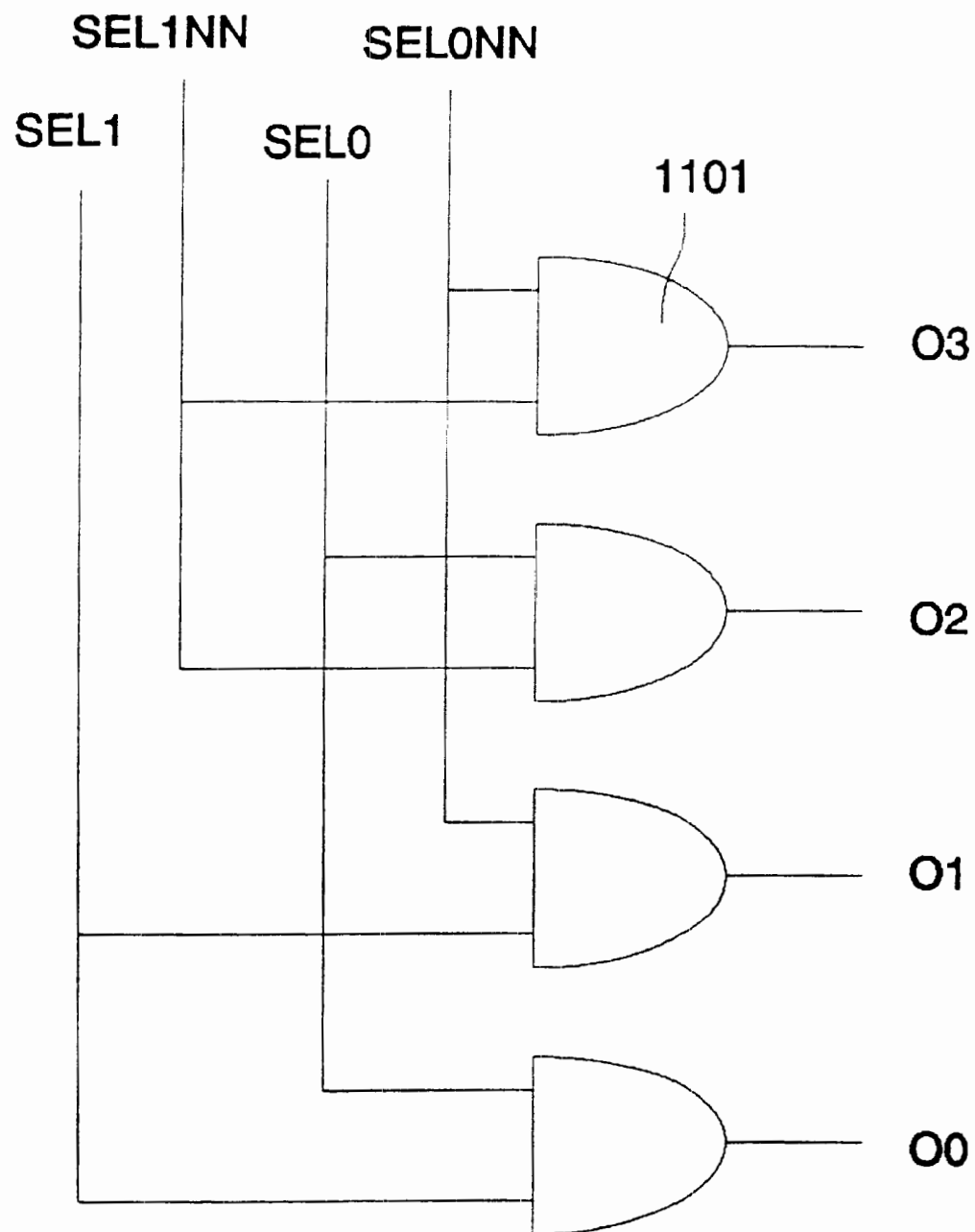


FIG. 11(a)

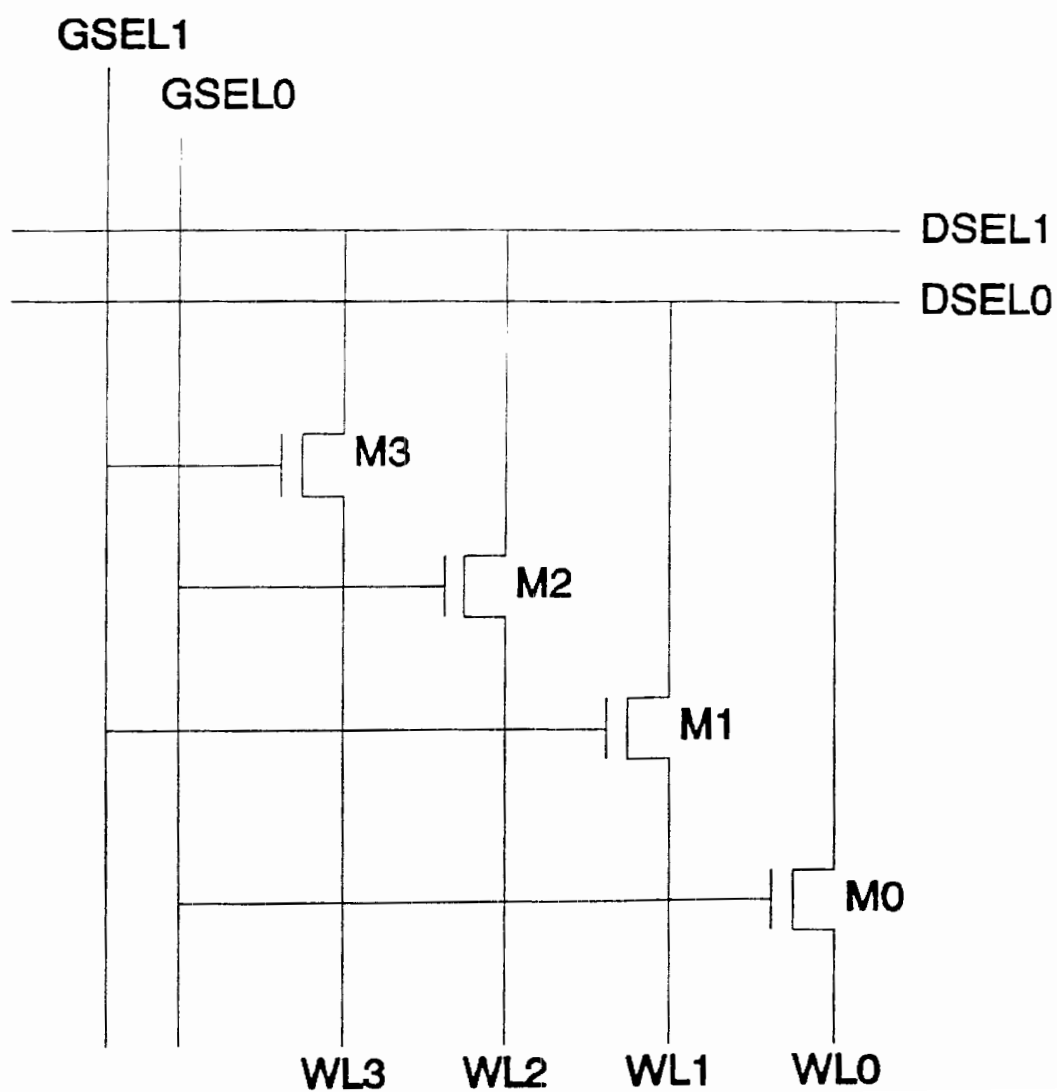


FIG. 11(b)

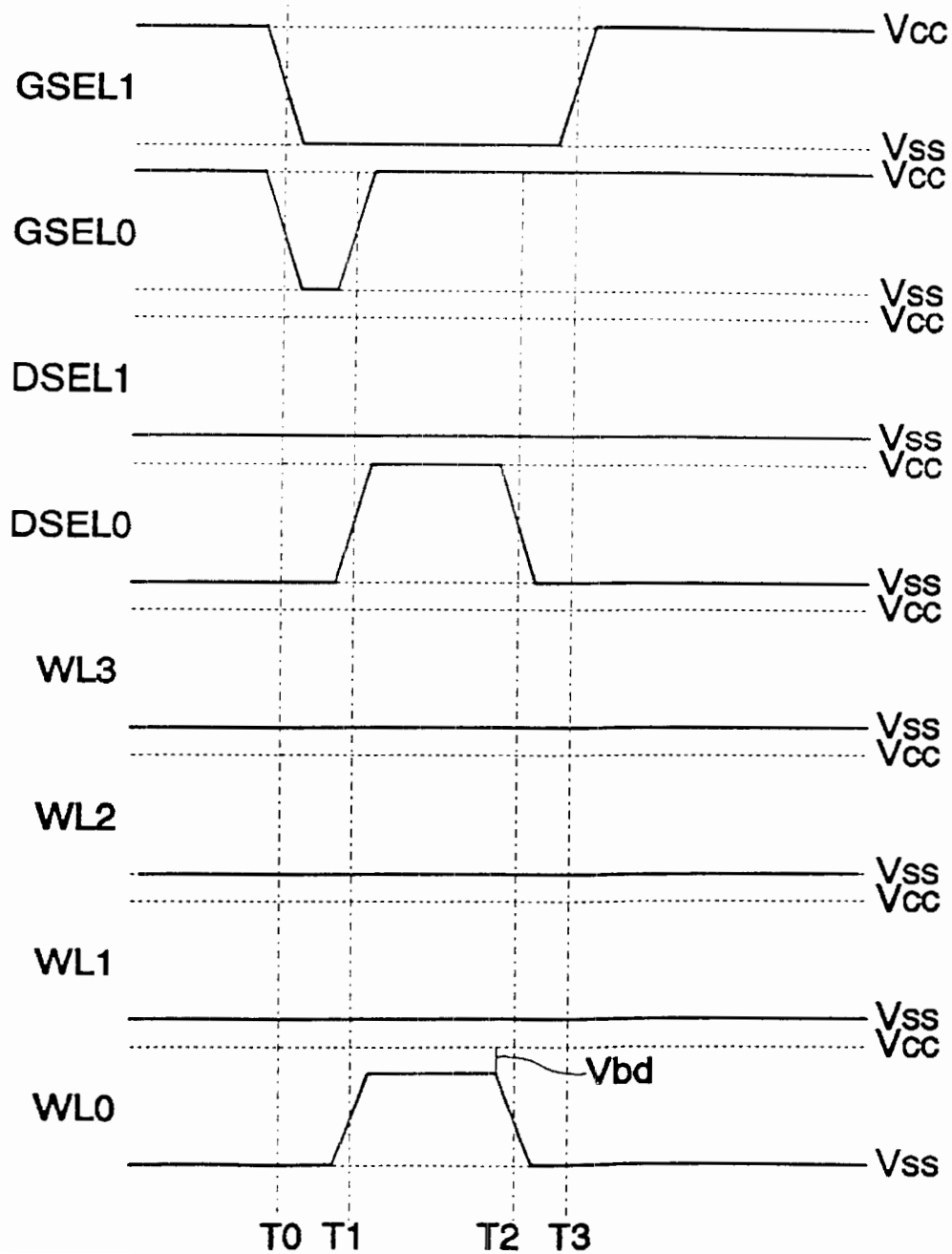


FIG. 12(a)

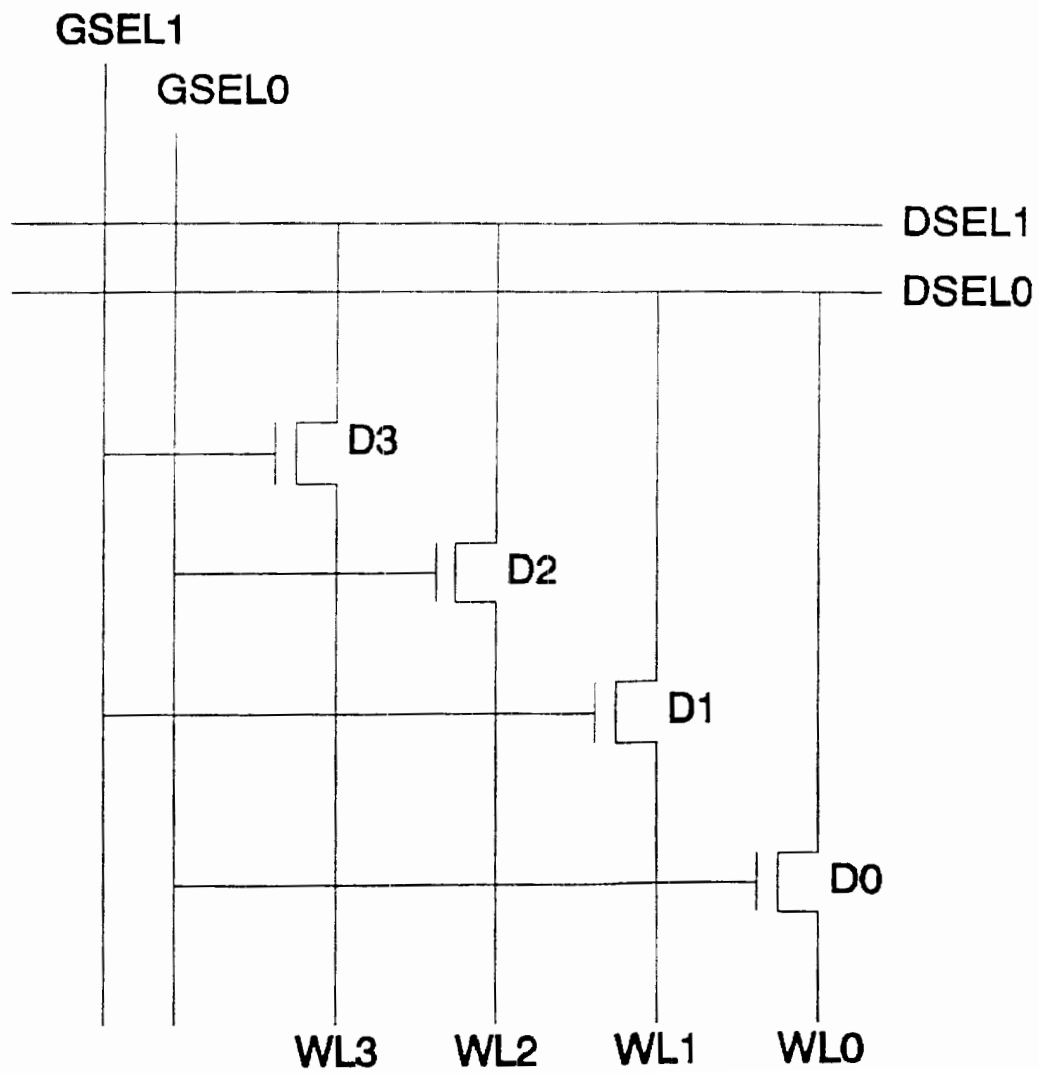


FIG. 12(b)

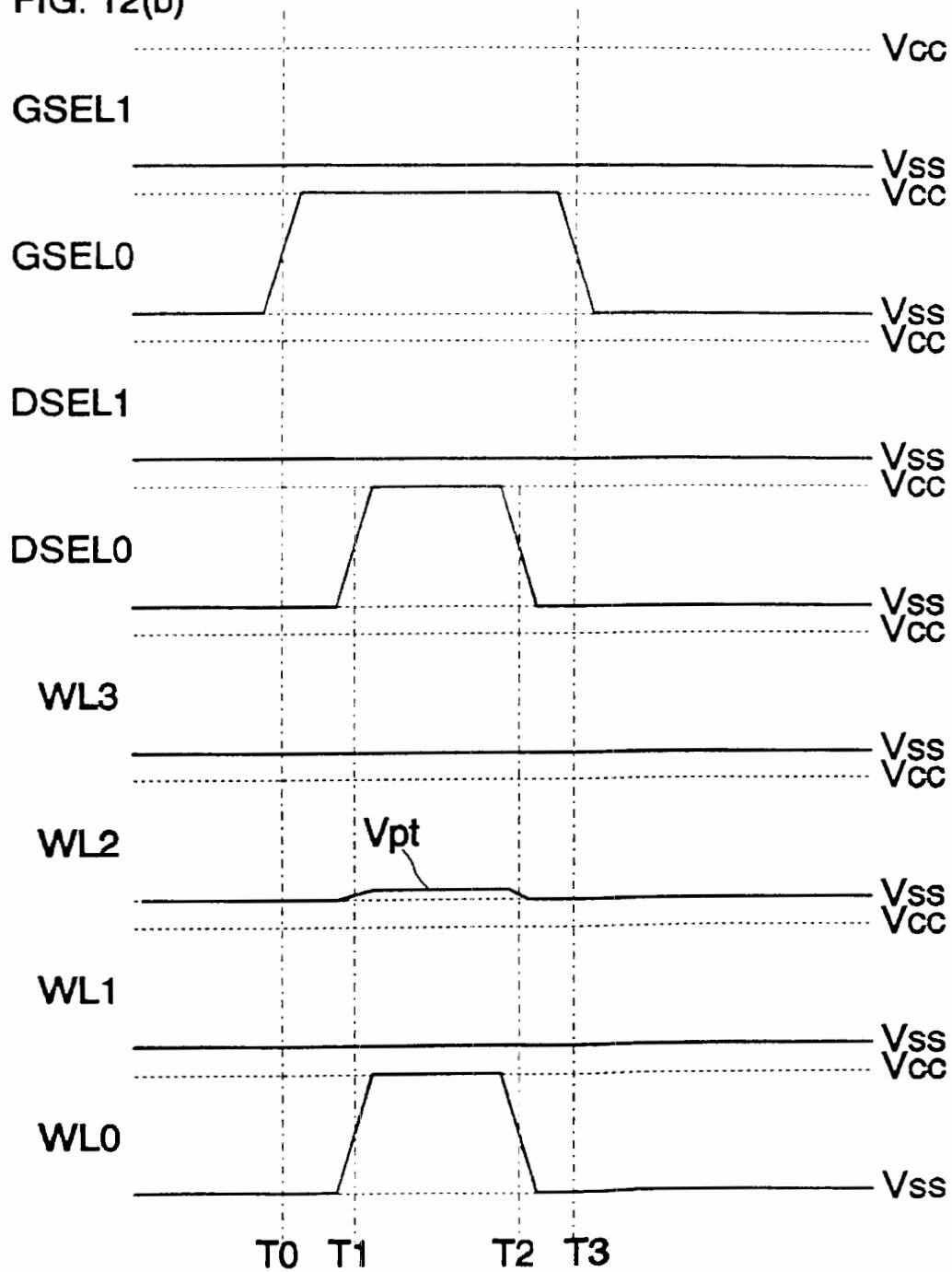
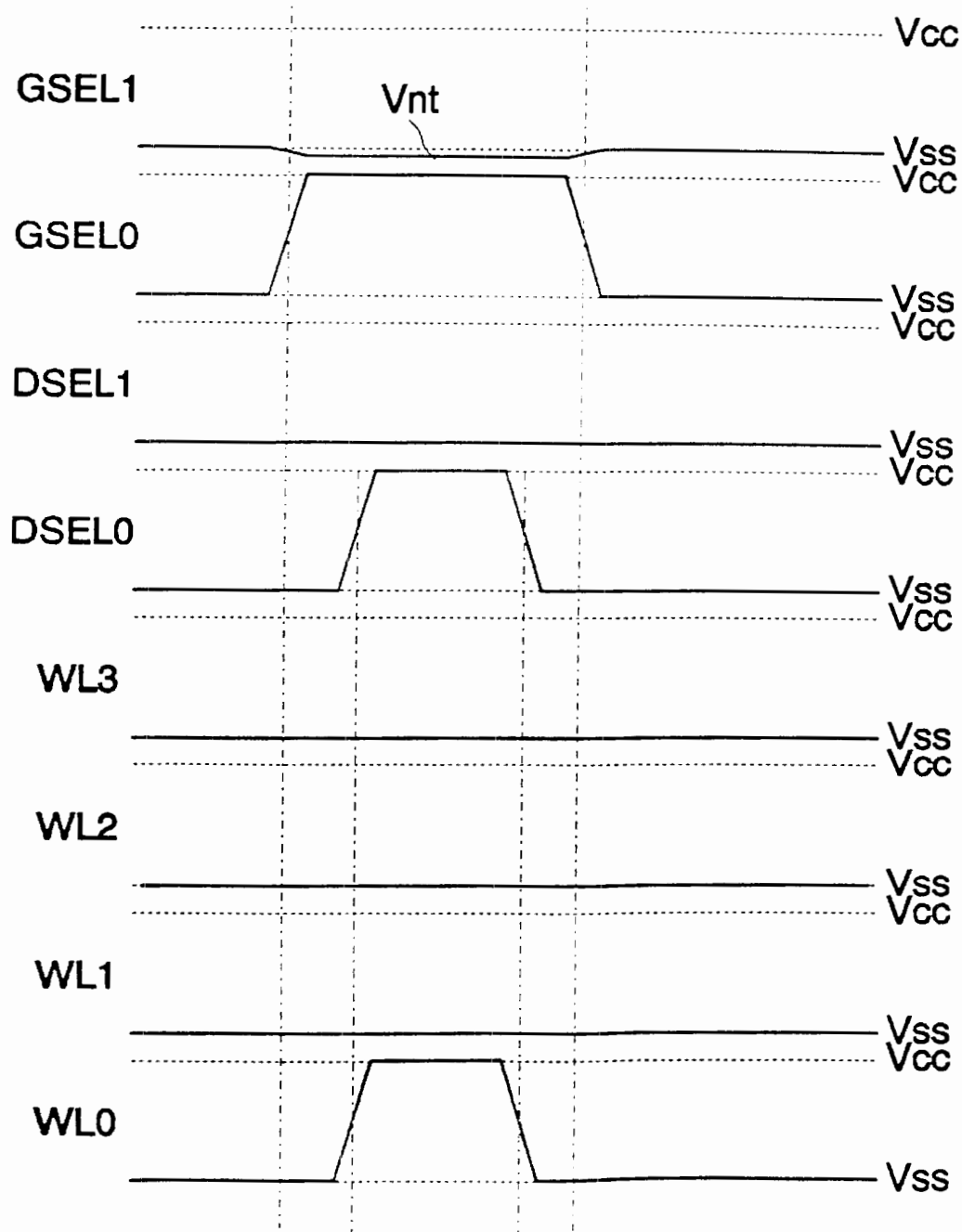


FIG. 12(c)





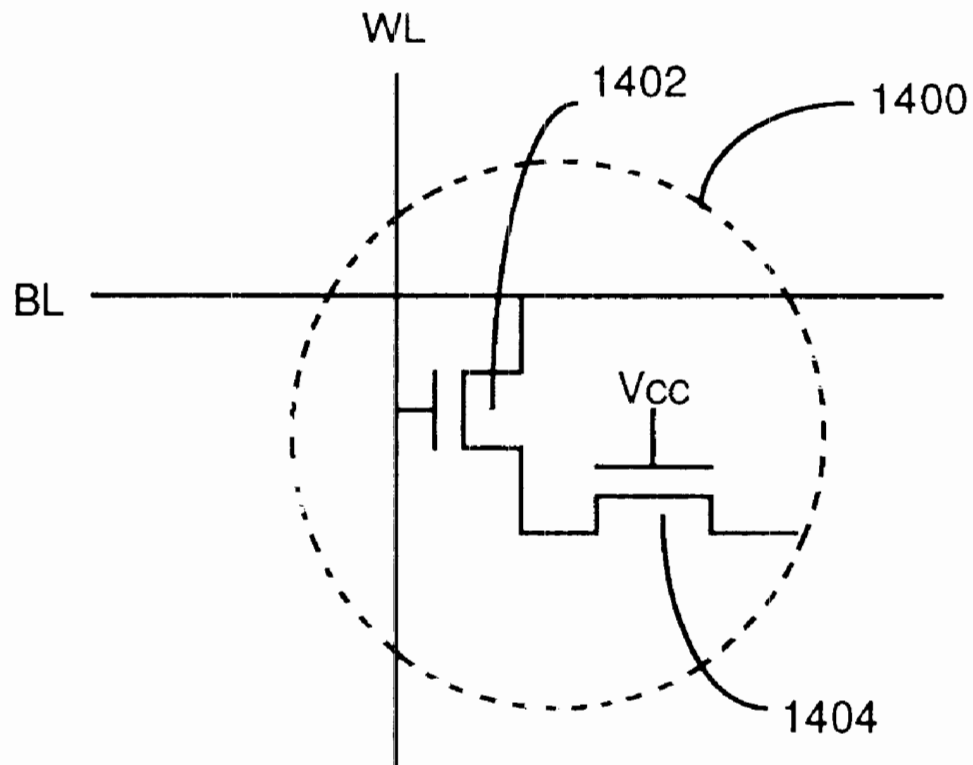


FIG. 13

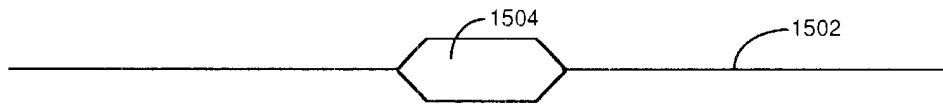


FIG. 14(a)

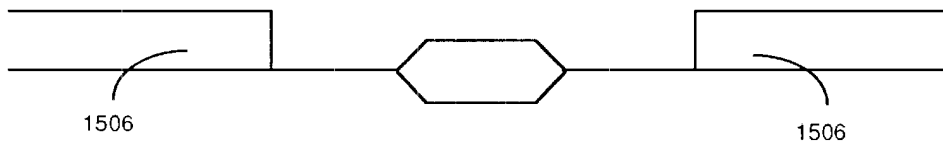


FIG. 14(b)

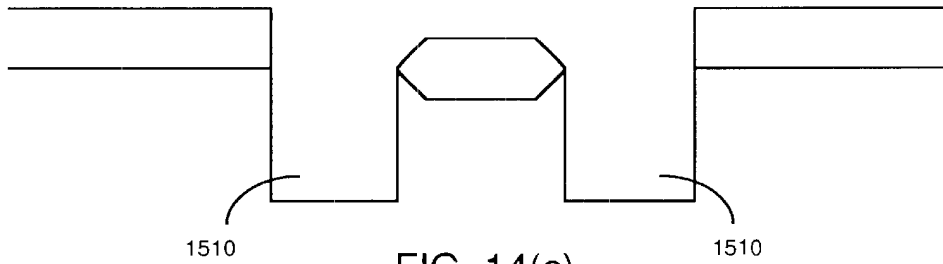


FIG. 14(c)

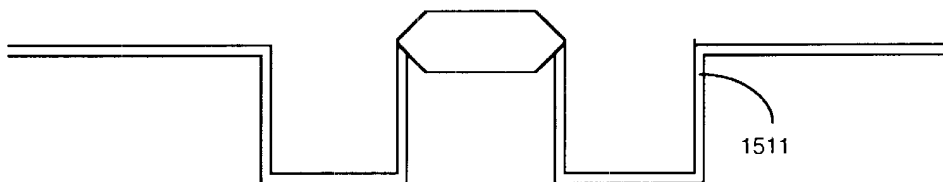


FIG. 14(d)

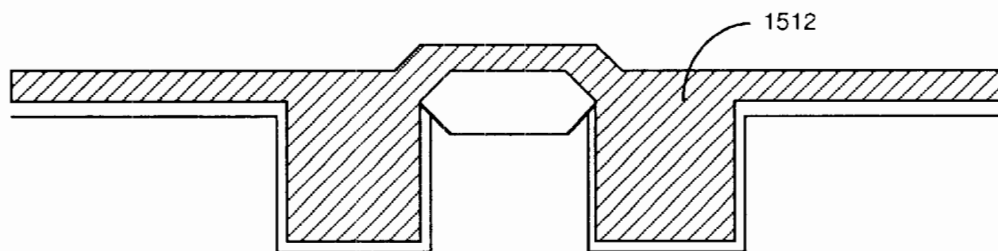


FIG. 14(e)

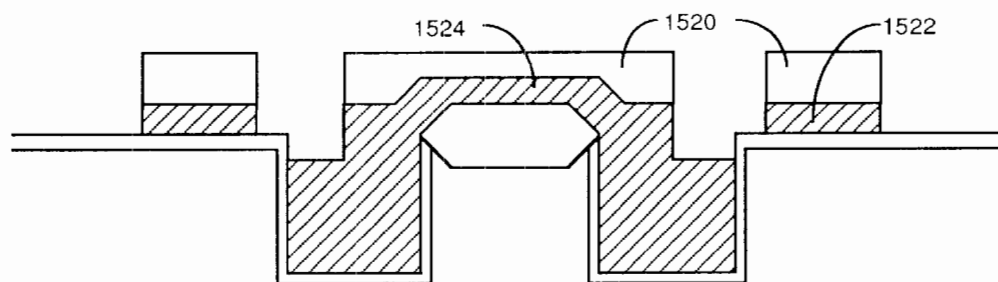


FIG. 14(f)

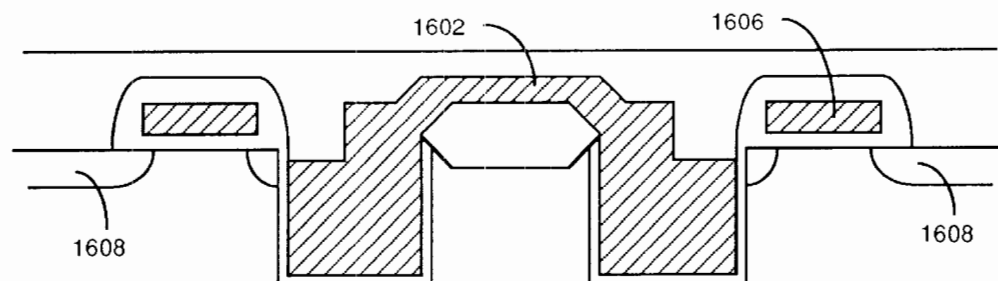


FIG. 14(g)

FIG. 15(a)

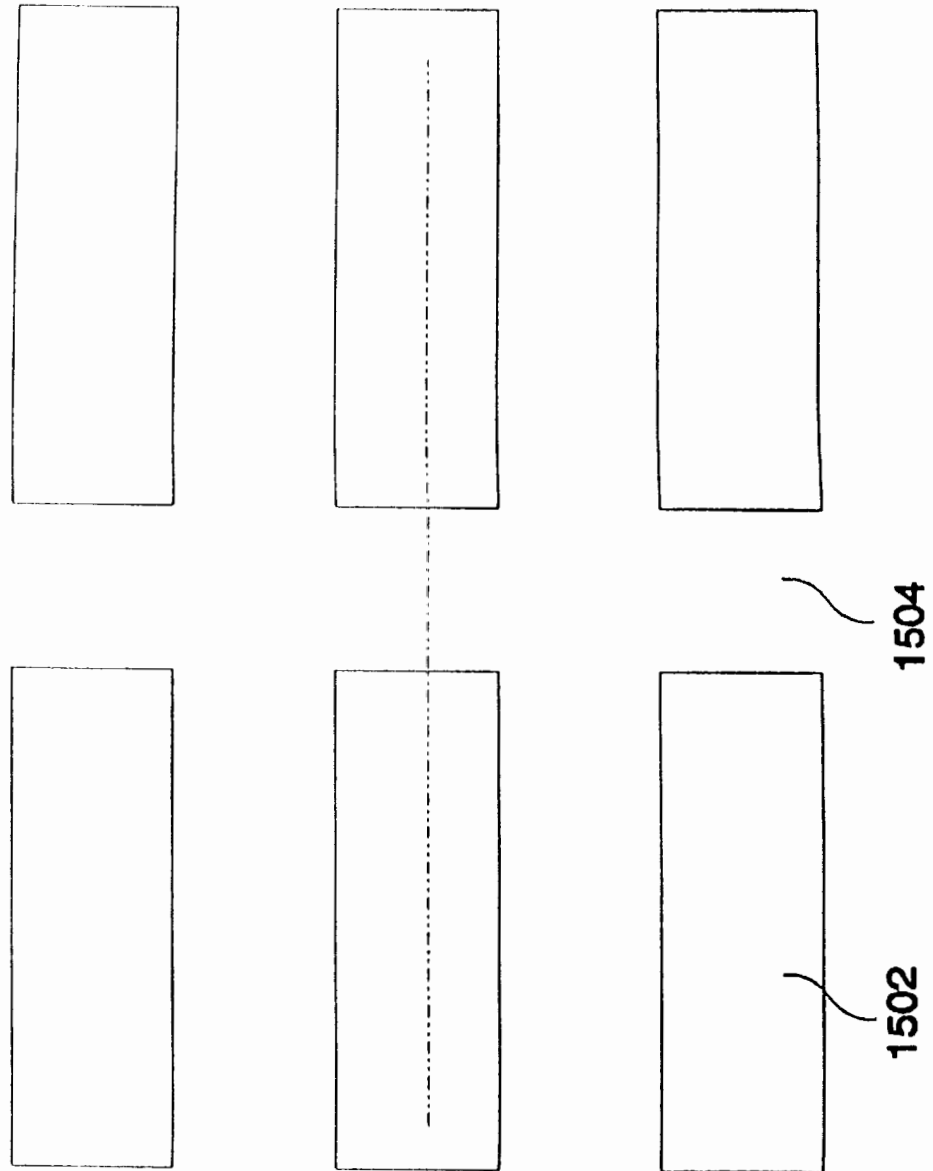


FIG. 15(b)

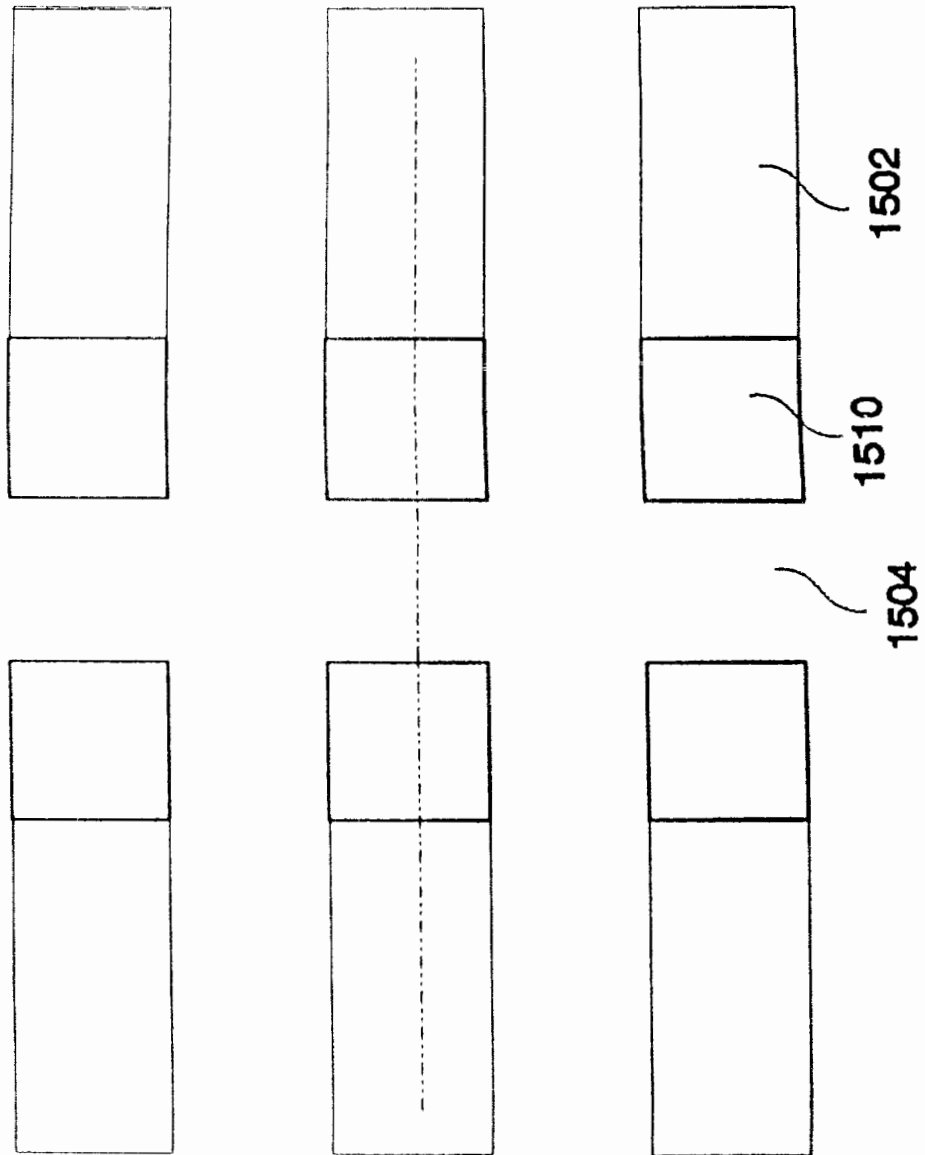
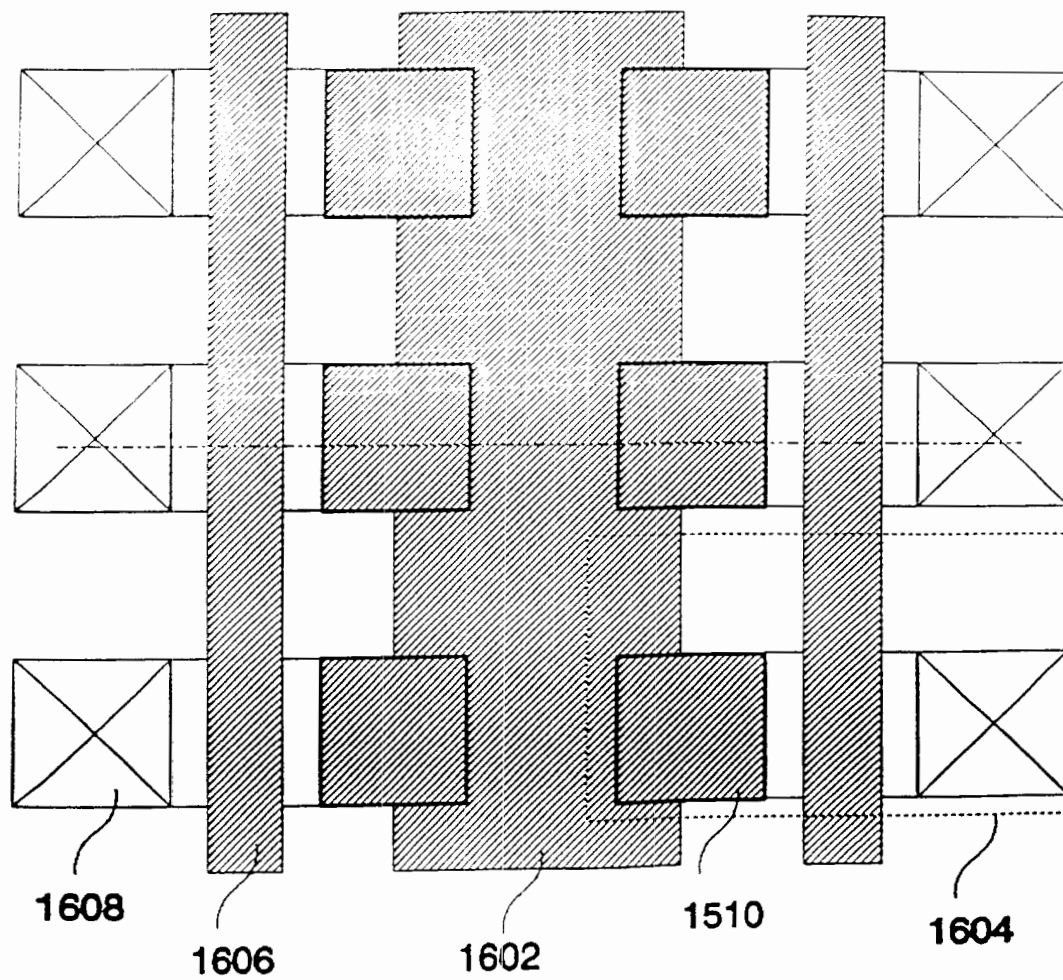


FIG. 15(c)





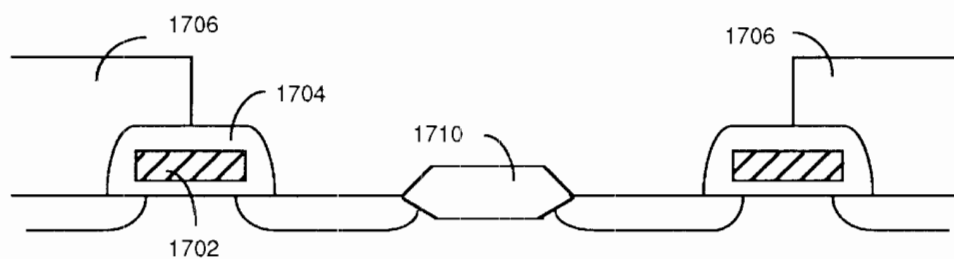


FIG. 16(a)

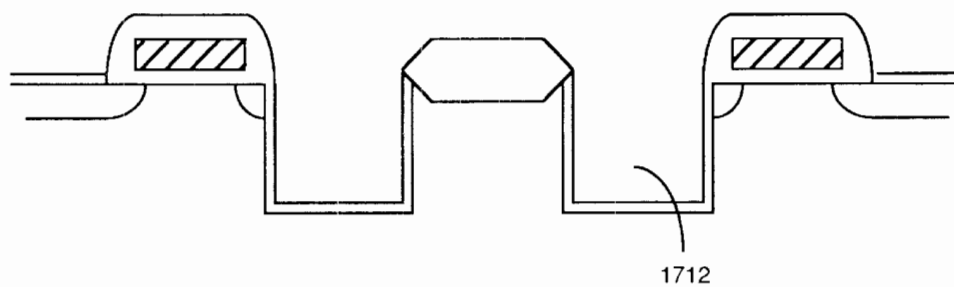


FIG. 16(b)

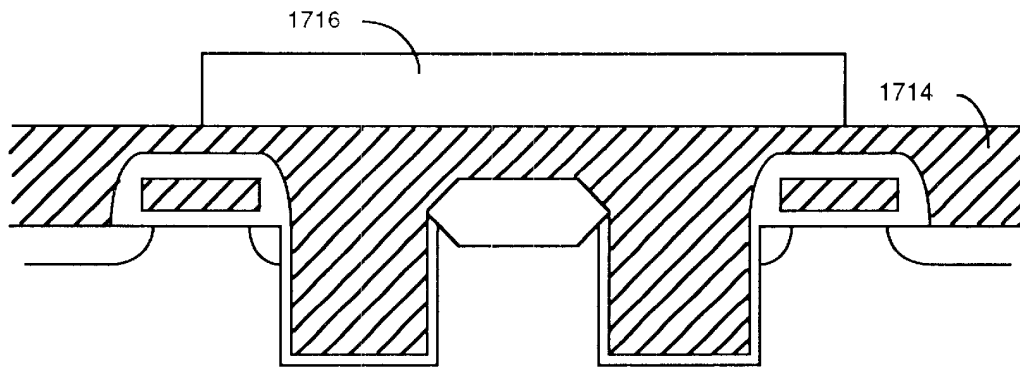


FIG. 16(c)

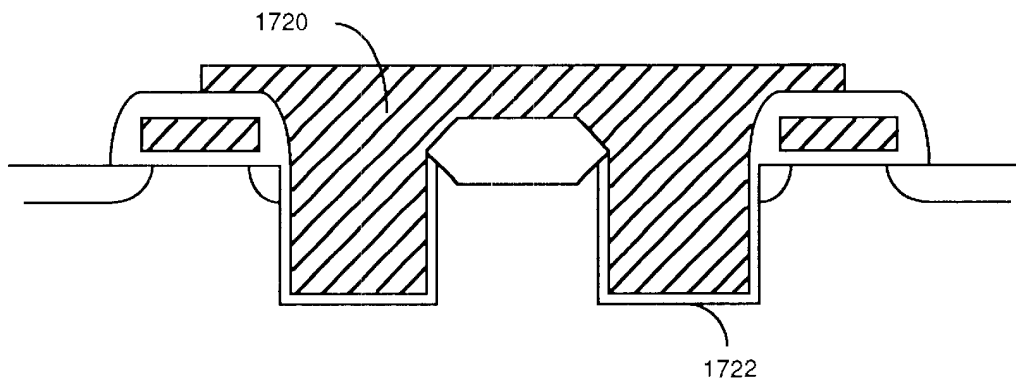
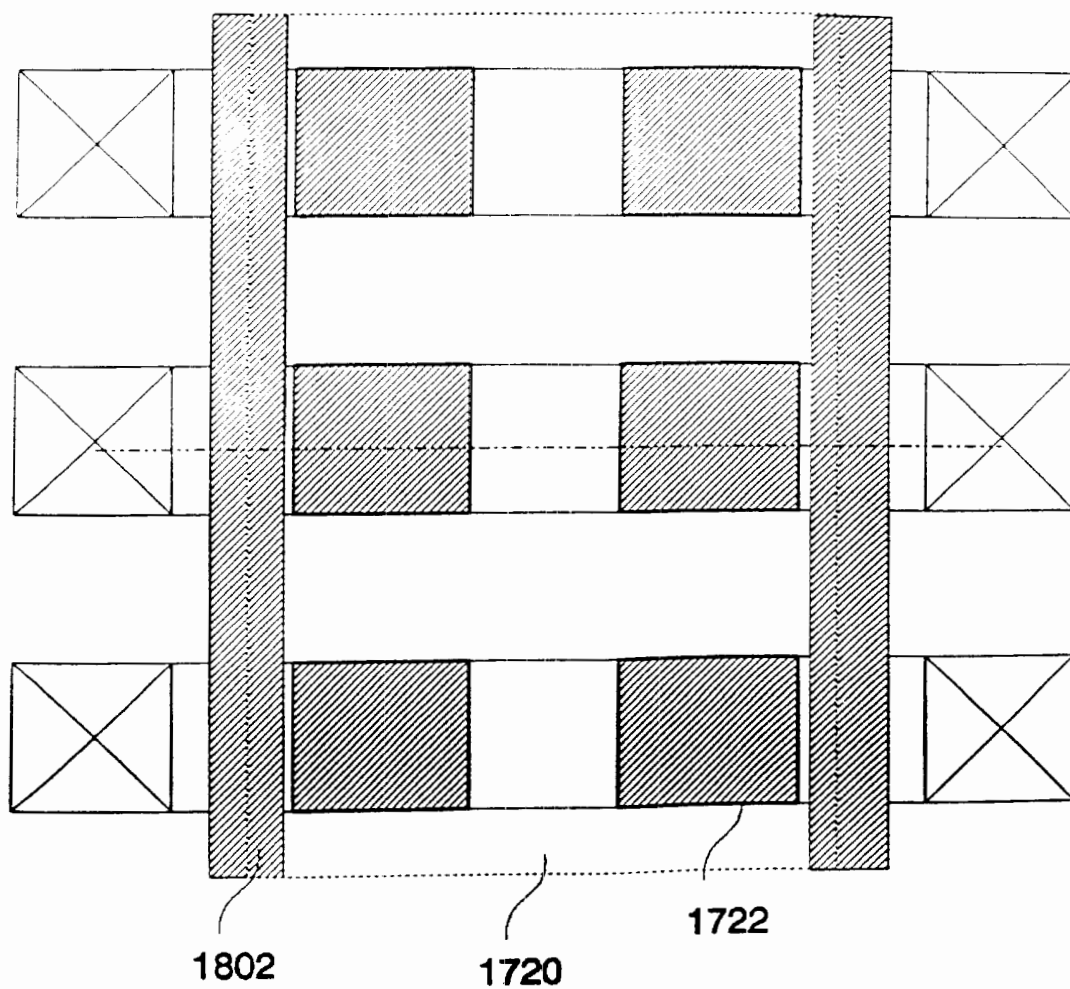


FIG. 16(d)

FIG. 17



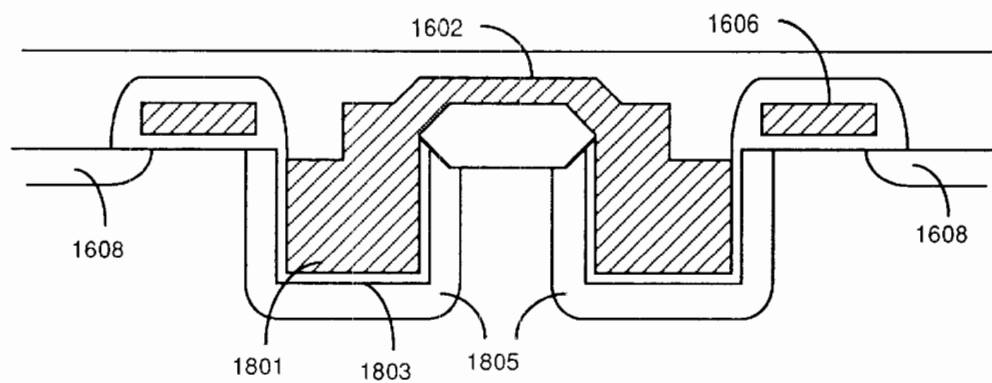


FIG. 18(a)

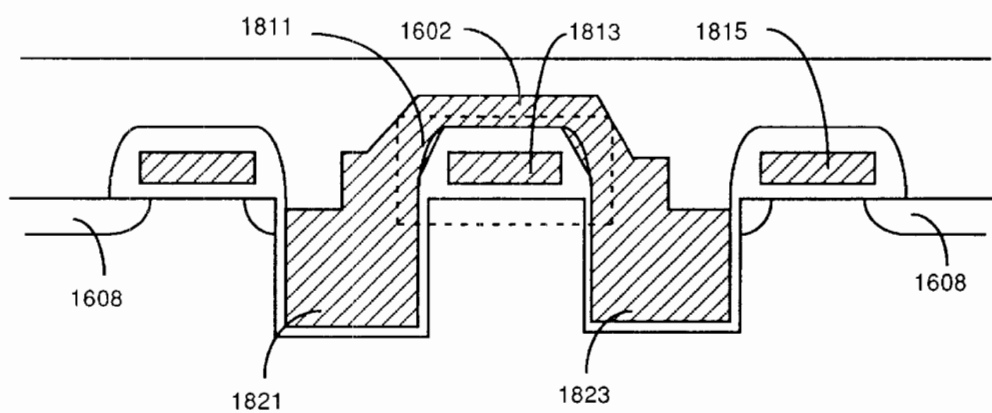
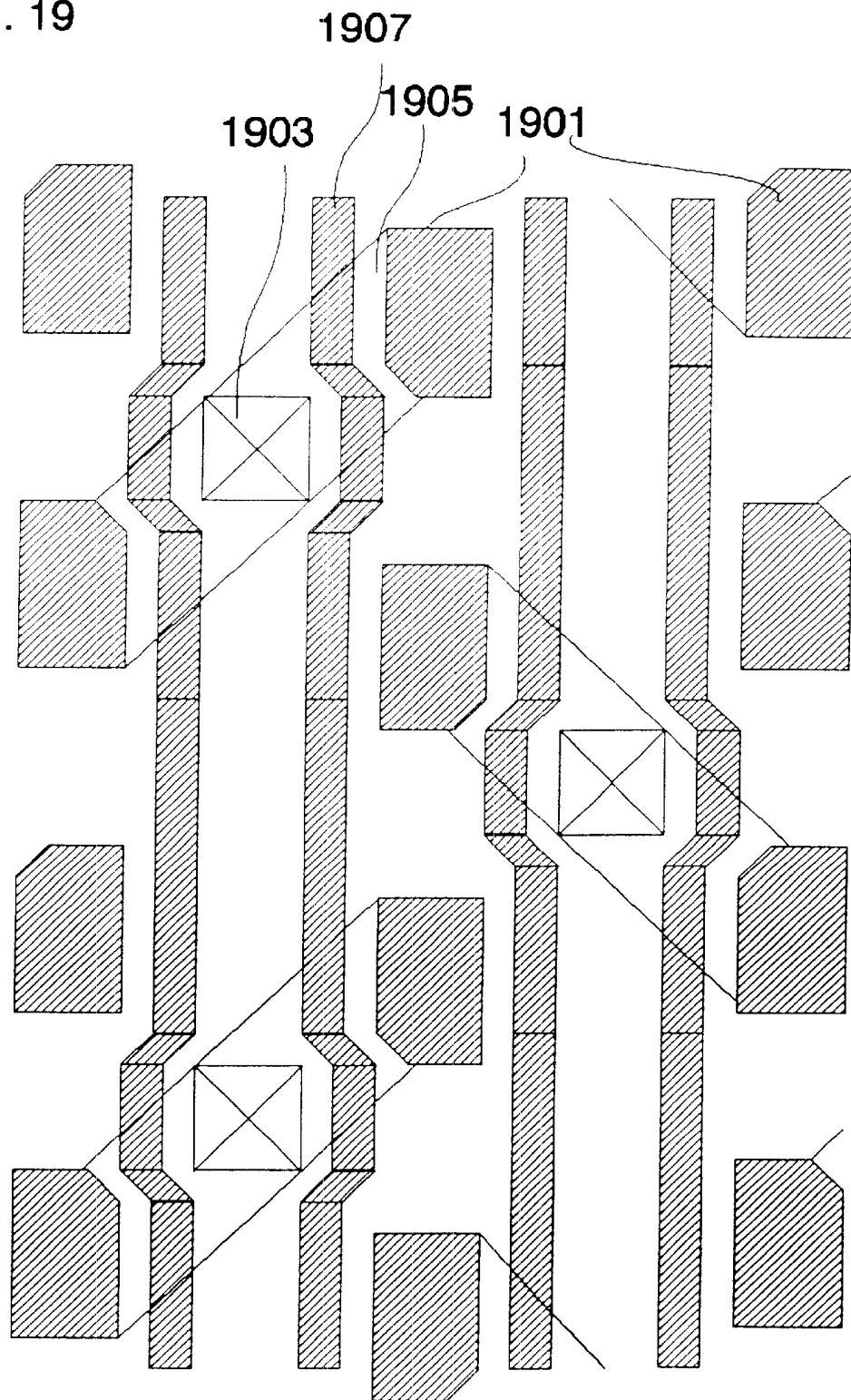
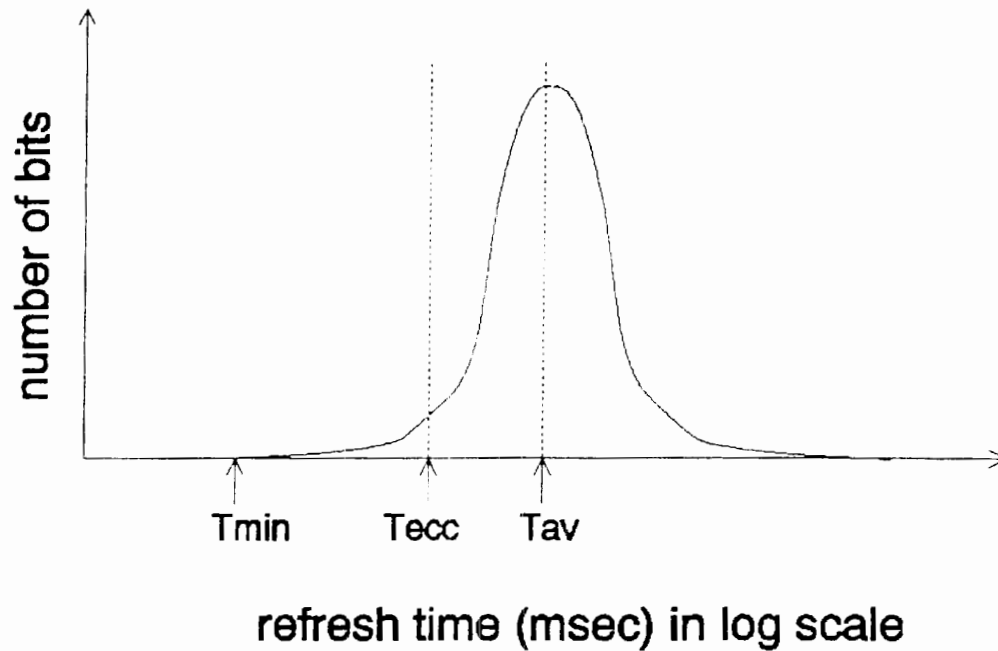
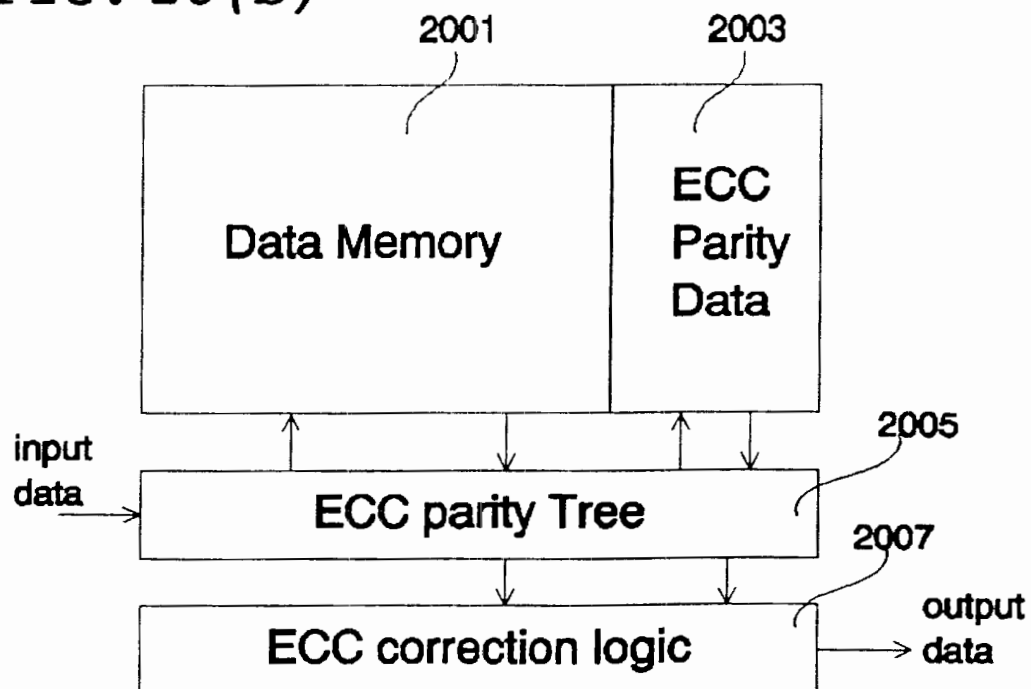


FIG. 18(b)

FIG. 19



*FIG. 20(a)**FIG. 20(b)*



US 6,687,148 B2

1

# **HIGH PERFORMANCE EMBEDDED SEMICONDUCTOR MEMORY DEVICES WITH MULTIPLE DIMENSION FIRST- LEVEL BIT-LINES**

The application Ser. No. 09/860,215 is a Continuation in Part (CIP) Application of application Ser. No. 08/653,620 filed on May 24, 1996 now U.S. Pat. No. 5,748,547 and another co-pending application Ser. No. 08/805,290 filed on Feb. 25, 1997 now U.S. Pat. No. 5,825,904 and an International Application filed in Taiwan Intellectual Property Bureau by identical sole inventor as for this CIP Application by identical sole inventor as for this Continuation-in-Part (CIP) Application.

## **BACKGROUND OF THE INVENTION**

### **1. Field of the Invention**

The present invention relates to high performance semiconductor memory devices, and more particularly to embedded memory devices having first level bit lines connected along different layout directions.

### **2. Description of the Prior Art**

DRAM is usually considered as a high density, low cost, but low performance memory device. DRAM's of current art always have lower performance relative to other types of semiconductor memories such as static random access memory (SRAM). The density of DRAM has been improved rapidly; the extent of integration has been more than doubled for every generation. Such higher integration of DRAM has been realized mainly by super fine processing technique and improvements in memory cell structure. In the mean time, the improvement in DRAM performance is progressing at a much slower rate. This relatively slower improvement rate in performance generates a performance gap between logic devices and memory devices. Many new approaches have been proposed to reduce this performance gap. The synchronized DRAM (SDRAM), the extended data output (EDO) DRAM, the multiple bank DRAM (MDRAM), and the RAMBUS system approaches are the most well known methods to improve DRAM performance. U.S. Pat. No. 4,833,653 issued to Mashiko et al. and U.S. Pat. No. 4,758,993 issued to Takemae et al. disclosed DRAM having selectively activated subarrays in order to improve performance. Another approach to improve DRAM performance is to place an SRAM cache into DRAM (called "hybrid memory"). U.S. Pat. No. 5,421,000 issued to Fortino et al., U.S. Pat. No. 5,226,147 issued to Fujishima et al., U.S. Pat. No. 5,305,280 issued to Hayano et al. disclosed embodiments of hybrid memories. The major problem for above approaches is that they are paying very high price for performance improvement, while the resulting memory performance improvement is still not enough to fill the gap. Another problem is that all of those approaches require special system design that is not compatible with existing computer systems; it is therefore more difficult to use them in existing computer systems.

Another disadvantage of DRAM is the need to refresh its memory. That is, the users need to read the content of memory cells and write the data back every now and then. The system support for DRAM is more complex than SRAM because of this memory refresh requirement. Memory refresh also represents a waste in power. U.S. Pat. No. 5,276,843 issued to Tillinghast et al. disclose a method to reduce the frequency of refresh cycles. U.S. Pat. No. 5,305,280 issued to Hayano et al. and U.S. Pat. No. 5,365,487 issued to Patel et al. disclosed DRAM's with self-

2

refresh capability. Those inventions partially reduce power consumption by refresh operations, but the magnitude of power saving is very far from what we can achieve by the present invention. The resource conflict problem between refresh and normal memory operations also remains unsolved by those patents.

Recently, Integrated Device Technology (IDT) announced that the company can make DRAM close to SRAM performance by cutting DRAM into small sub-arrays. The new device is not compatible with existing memory; it requires special system supports to handle conflicts between memory read operation and memories refresh operation. It requires 30% more area the DRAM, and its performance is still worse than SRAM of the same size.

Another important problem for DRAM design is the tight pitch layout problem of its peripheral circuits. In the course of the rapid improvement in reducing the size of memory cells, there has been no substantial improvement or change as to peripheral circuits. Peripheral circuits such as sense amplifiers, decoders, and precharge circuits are depend upon memory cell pitch. When the memory cells are smaller for every new generation of technology, it is more and more difficult to "squeeze" peripheral circuits into small pitch of memory layout. This problem has been magnified when the memory array is cut into smaller sub-arrays to improve performance. Each subarray requires its own peripheral circuits; the area occupied by peripheral circuits increases significantly. Therefore, in the foreseeable future, there may occur a case wherein the extent of integration of DRAM is defined by peripheral circuits. U.S. Pat. No. 4,920,517 issued to Yamauchi et al. disclosed a method to double the layout pitch by placing sense amplifiers to both ends of the memory. This method requires additional sense amplifiers. Although the available layout pitch is wider than conventional DRAM, the layout pitch is still very small using Yamauchi's approach.

All of the above inventions and developments provided partial solutions to memory design problems, but they also introduced new problems. It is therefore highly desirable to provide solutions that can improve memory performance without significant degradation in other properties such as area and user-friendly system support.

Another difficulty encountered by those of ordinary skill in the art is a limitation that Dynamic Random Access Memory (DRAM) which is usually considered as a high density, low cost, and low performance memory device cannot be conveniently integrated as embedded memory. This is due to the fact that higher integration of DRAM has been realized mainly by super fine processing technique and improvements in memory cell structure. A typical DRAM manufacture technology of current art is the four layer poly silicon, double layer metal (4P2M) process. Such memory technology emphasizes on super-fine structure in manufacture memory cells; performance of it logic circuit is considered less important. A technology optimized to manufacture high speed logic products have completely different priority; it emphasizes on performance of transistors, and properties of multiple layer metals. An example of a typical logic technology of current art is the triple layer metal, single poly silicon (1P3M) technology.

An embedded memory, by definition, is a high density memory device placed on the same chip as high performance logic circuits. The major challenge to manufacture high density embedded memory is the difficulty in integrating two types of contradicting manufacture technologies together. An embedded technology of current art requires 4

US 6,687,148 B2

3

layers of poly silicon and 3 layers of metal. There are more than 20 masking steps required for such technology. It is extremely difficult to have reasonable yield and reliability from such complex technology of current art. Further more, the current art embedded technology tend to have poor performance due to contradicting requirements between logic circuits and memory devices. None of current art embedded memory technology is proven successful. There is an urgent need in the Integrated Circuit (IC) industry to develop successful embedded memory devices.

The Applicant of this Patent Application has been successful in manufacturing embedded memory devices by novel approaches to change the architecture of IC memory so that the memory device no longer has conflicting properties with logic circuits. Examples of such architecture change have been disclosed in co-pending patent application Ser. No. 08/653,620. The previous application solved the tight pitch layout problems along the sense amplifier location, and it solves the self-refresh requirement by hiding refresh requirements. This CIP Application further discloses solutions for remaining problems. A single-transistor decoder circuit solves the tight pitch layout problem along the decoder direction. Typical logic technology or small modification of existing logic technology may be applied to manufacture the memory cells. Using these novel inventions, high performance and high density embedded memory devices are ready to be manufactured.

#### SUMMARY OF THE PRESENT INVENTION

The primary objective of this invention is, therefore, to improve the performance of semiconductor memory device without paying extensive area penalty. Another primary objective is to make DRAM more user-friendly by making the performance improvement in parallel with simplification in system supports. Another primary objective is to provide an improved semiconductor memory device in which peripheral circuits can readily follow further higher integration of memory cells. Another objective is to reduce power consumption of high performance semiconductor memory.

Another important objective of this invention is to manufacture high-density memory device on the same chip with high performance logic devices without using complex manufacture technology. Another primary objective is to make embedded DRAM to have the same performance as high-speed logic circuits. Another primary objective is to improve yield and reliability of embedded memory products.

These and other objects are accomplished by a semiconductor memory device according to the invention. The memory device includes a novel architecture in connecting bit lines along multiple layout directions, a new design in decoder circuit, and a novel timing control that can finish a read cycle without waiting for completion of memory refresh.

According to the present invention as described herein, the following benefits, among others, are obtained.

- (1) The multiple dimensional bit line structure dramatically reduces the parasitic loading of bit lines seen by sense amplifiers. Therefore, we can achieve significant performance improvement. Our results show that a memory of the present invention is faster than an SRAM of the same memory capacity.
- (2) The multiple dimension bit line structure also allows us to use one sense amplifier to support many bit line pairs. Therefore, we no longer have tight pitch layout problem for sense amplifiers and other peripheral circuits. Remov-

4

ing tight pitch problem allows us to achieve performance improvement without paying high price in layout area.

- (3) A novel decoder design reduces the size of memory decoder dramatically, that allow designers to divide the memory array into sub-arrays without paying high price in the area occupied by decoders.
- (4) A novel input and output (IO) circuit design allows us to delay the memory refresh procedures until next memory operation. This approach allows us to "hide" refresh cycles and memory update cycles in a normal memory operation. The resulting memory device is as friendly as existing SRAM device. In fact, a memory of this invention can be made fully compatible with existing SRAM device.
- (5) All of the above improvements are achieved by using much lower power than the power used by prior art DRAM's.
- (6) The tight pitch layout problem along the decoder direction is solved. Therefore, we can divide a memory array into smaller blocks without sacrificing significant area. This architecture change allows us to use smaller storage capacitor for each DRAM memory cell, which simplifies manufacture procedure significantly.
- (7) High density DRAM memory cells can be manufacture by adding simple processing steps to logic IC technology of current art. The resulting product supports high performance operation for both the memory devices and the logic circuits on the same chip.
- (8) The simplification in manufacturing process results in significant improvements in product reliability and cost efficiency.

While the novel features of the invention are set forth with particularly in the appended claims, the invention, both as to organization and content, will be better understood and appreciated, along with other objects and features thereof, from the following detailed description taken in conjunction with the drawing.

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a schematic block diagram of a prior art memory device;

FIG. 2 is a simplified block diagram of a multiple bank semiconductor memory device;

FIG. 3a is a schematic block diagram of a memory device with two-dimensional bit lines;

FIG. 3b is a schematic block diagram of a memory device with three-dimensional bit lines;

FIG. 4a is an illustration showing layout and power consumption of a prior art memory bank;

FIG. 4b is an illustration showing layout and power consumption of a semiconductor memory device of a first embodiment according to the invention;

FIG. 5 is a schematic diagram of the sense amplifier used by this invention;

FIG. 6 is a schematic diagram of the IO circuits of the present invention;

FIG. 7a shows the waveforms of critical signals during a read cycle;

FIG. 7b shows the waveforms of critical signals during a refresh cycle;

FIG. 7c shows the waveforms of critical signals during a write cycle;

FIG. 8 is a schematic diagram of the IO circuits of the present invention to support faster data read; and

FIG. 9 shows the timing relationship of critical signals of a memory device of this invention.

US 6,687,148 B2

5

FIG. 10 shows an example of a prior art CMOS decoder;

FIG. 11(a) is a schematic diagram of an enhance mode single transistor decoder of the present invention, and FIG. 11(b) is a diagram for the control signals and output signals of the decoder in FIG. 11(a);

FIG. 12(a) is a schematic diagram of a depletion mode single transistor decoder of the present invention, and FIGS. 12(b,c) illustrate the control signals and output signals of the decoder in FIG. 12(a);

FIG. 13 is a schematic diagram of a memory cell that uses an active transistor device as the storage capacitor of the memory cell;

FIGS. 14(a–g) are cross-section diagrams describing the process step to manufacture a DRAM memory cell by adding one masking step to standard logic technology;

FIGS. 15(a–c) are top views of the process step to manufacture a DRAM memory cell by adding one masking step to standard logic technology;

FIGS. 16(a–d) are cross-section diagrams describing another process step to manufacture a self-aligned trench capacitor for DRAM memory cell using one additional mask to standard logic technology;

FIG. 17 shows the top view of the memory cell manufactured by the process illustrated in FIGS. 16(a)–(d);

FIG. 18(a) shows the cross-section structures for capacitors that do not have the electrode voltage polarity constraint;

FIG. 18(b) shows the cross-section structures for memory cells that use transistors to separate nearby trench capacitors;

FIG. 19 illustrates the top view structure of practical memory cells of the present invention;

FIG. 20(a) shows a typical distribution of memory refresh time for the memory cells in a large DRAM; and

FIG. 20(b) is a symbolic diagram for a DRAM equipped with error-correction-code (ECC) protection circuit.

#### DETAILED DESCRIPTION OF THE INVENTION

Before the invention itself is explained, a prior art semiconductor memory-device is first explained to facilitate the understanding of the invention.

FIG. 1 shows memory cell array structure of a prior art DRAM in both electrical and topographical manners. Memory cell array 100 includes plural pairs of bit lines BL1, BL1#, BL2, BL2#, BL3, BL3#, . . . ; BLn, BLn# (n; integer) which are disposed in parallel manner and a plurality of word lines WL1, WL2 . . . WLn (m; integer) which are disposed in parallel manner and also in such manner that they intersect with bit lines perpendicularly. At intersecting points, memory cells MC1, MC2, . . . , MCn are disposed. Memory cells are shown by circle marks in memory cell array 100 in FIG. 1. Each memory cell includes a switching field effect transistor 110 and memory cell capacitor 112. Bit line BL is connected to the drain of the transistor 110. The gate of transistor 110 is connected to word line WL. Sense amplifiers SA1, SA2, . . . SAn are disposed at one end of memory cell array and each pair of bit lines are connected to one sense amplifier. For example, a pair of bit lines BL1, BL1# are connected to sense amplifier SA1, a pair of bit lines BL2, BL2# are connected to sense amplifier SA2 . . . , and a pair of bit lines BLn, BLn# are connected to sense amplifier SAn. The outputs of those sense amplifiers are connected to data output switches 120. The output switches 120 contain a multiplexer 122 that is controlled by a decoder

6

124. The output switches 120 select the outputs from one of the sense amplifiers, and place the data on the data buses D and D#.

For example, when information is read out from memory cell MC1, the following operations are carried out. First, word line WL2 is selected by the word line decoder 126 and the transistor 110 in memory cell MC1 is rendered conductive. Thereby, signal charge in capacitor 112 of memory cell MC1 is read out to bit line BL1# so that minute difference of electric potential occurs between a pair of bit lines BL1 and BL1#. The sense amplifier SA1 amplifies such difference. The output switches 120 select the outputs of SA1 and thereafter, transfer the data to data buses D, D# through a multiplexer 122. After the above read procedure, the charge stored in the cell capacitor 112 is neutralized. It is therefore necessary to write the original data sensed by SA1 back to the memory cell MC1. Such procedure is called “refresh”. The sense amplifier used in current art always refreshes the memory cell after it determines the state of the memory cell. It is very important to remember that all the other memory cells along the word line, MC2, MC3, . . . MCn, are also rendered conductive when WL2 is selected. It is therefore necessary to turn on all the other sense amplifiers SA2, SA3, . . . SAn to read and refresh the data stored in all other memory cells connected to WL2, when we only need the data stored in MC1.

DRAM of such structure has the following drawbacks.

- (1) In order to read the data from a few memory cells along one word line, we need to read and refresh all the memory cells along that word line. Most of the energy is used for refreshing instead of reading data. This waste in energy also results in slower speed because a large number of devices need to be activated.
- (2) As the size of the memory array increases, the bit line parasitic capacitance (Cb) increases. The ratio between the memory cell capacitance Cm and the bit line parasitic capacitance Cb determines the amplitude of the potential difference on the bit line pairs. The memory read operation is not reliable if the (Cm/Cb) ratio is too small. Thereby, the (Cm/Cb) ratio is often the limiting factor to determine the maximum size of a memory array. Special manufacturing technologies, such as the trench technology or the 4-layer poly technology, have been developed to improve the memory cell capacitance Cm. However, the Cm/Cb ratio remains a major memory design problem.
- (3) To support refresh procedures, we always need to have one sense amplifier for each bit line pair. As higher integration of memory cells progresses, the layout pitch for sense amplifier decreases. Thereby, it becomes difficult to form stable and well operable sense amplifier within the pitch. Such problem is often referred as the “tight pitch layout” problem in the art of integrated circuit design. Tight pitch layout always results in excessive waste in silicon area due to the difficulty in squeezing active devices into a narrow space. Similar problem applies to other peripheral circuits such as decoders and pre-charge circuits.

To reduce the effect of the above problems, large memory of prior art is always divided into plural sub-arrays called memory banks 200 as shown in FIG. 2. Each bank 200 of the memory sub-array has its own decoder 210 and output switches 212. Each pair of the bit lines in each memory bank needs to have one sense amplifier 214. The outputs of each memory bank are selected by output switches 212, and placed on data buses 220 so that higher order amplifiers and decoders can bring the data to output pins.



This multi-bank approach provides partial solutions to the problems. Because each memory bank is capable of independent operation, we can reduce power consumption by keeping unused memory banks in low power state. The speed is also improved due to smaller active area. The (Cm/Cb) ratio can be kept at proper value by limiting the size of each memory bank. Multiple-bank memory allows us to turn on a sub-set of sense amplifiers to save power, but each bit line pair still needs to have one sense amplifier because we still need to refresh the contents of all activated memory cells. This multi-bank approach provides partial solutions, but it creates new problems. Each memory bank needs to have a full set of peripheral circuits; the areas occupied by the peripheral circuits increase significantly. Smaller size of memory bank implies higher percentage of area spent on peripheral circuits. Balancing the requirement between (Cm/Cb) ratio and the increase in tight pitch layout peripheral circuits is a major design problem for multiple bank memories. Yamauchi et al. were able to double the pitch for sense amplifiers by placing sense amplifiers at both sides of the memory array, but the layout pitch is still too small. Many other approaches have been proposed, but all of them provided partial solutions to part of the problems while created new problems.

This invention is made to solve the above-stated problems. FIG. 3a shows memory structure of one embodiment of the present invention in both electrical and topographical manners. The building block of the present invention is a memory unit 300. Each memory unit contains decoders 302, amplifiers AMP1, AMP2, . . . , AMPi, and a plurality of memory blocks 310. These memory blocks are arranged in pairs; memory block 1# is symmetrical to memory block 1; memory block 2# is symmetrical to memory block 2; . . . ; and memory block i# is symmetrical to memory block i. Each memory block contains word line switches 312, bit line switches 314, and a small memory array 316. The word line switches 312 and bit line switches 314 are controlled by block select signals. For example, the block select signal BLKSEL1 controls the word line switches and the bit line switches in memory block 1 and in memory block 1#. The memory array contains memory cells similar to the memory cells in FIG. 1. Circle marks are used to represent those memory cells in FIG. 3a. Each memory cell is connected to a short word line and a short bit line within each memory block. For example, in memory block 1 the gate of the memory cell MC12 is connected to block word line WL12 and block bit line BL12. Each block word line is connected to one unit word line through a word line switch 312. For example, WL12 is connected to UWL2 through a word line switch 312 controlled by block select signal BLKSEL1; WL22 is connected to UWL2 through a word line switch controlled by block select signal BLKSEL2; . . . ; WLij is connected to UWLj through a word line switch controlled by block select BLKSELi (i and j are integers). In this example, the memory unit has two levels of bit lines—the unit level bit lines UBL1, UBL1#, UBL2, UBL2#, . . . UBLn, UBLn# and the block level bit lines BL11, BL11#, BL12, BL12#, . . . et al. The block bit lines are made by the first layer metal (metal 1), and they are disposed vertical to the word lines. The unit bit lines are made by the second layer metal (metal 2), and they are disposed in parallel to the word lines. Each block bit line is connected to one unit bit line through one bit line switch 314 in each block. For example, BL12 is connected to UBL2 through a bit line switch controlled by block select signal BLKSEL1; BL22 is connected to UBL2 through a bit line switch also controlled by block select signal BLKSEL2; . . . ; BLii is connected to UBLi through

a bit line switch controlled by block select BLKSELi. Each pair of unit bit lines is connected to one amplifier. For example, UBL1 and UBL1# are connected to AMP1; UBL2 and UBL2# are connected to AMP2; . . . ; UBL1 and UBL1# are connected to AMPi. Those unit-bit-lines and block-bit-lines form a two-dimensional network that allows one amplifier to support bit line pairs in many blocks.

This two-dimensional bit line connection allows us to read the memory content with little waste in power. For example, when information is read out from memory cells on WL12 in block 1, the following operations are carried out. First, the block-select signal BLKSEL1 is activated, while all other block select signals remain inactive. All the word line switches 312 and bit line switches 314 in memory block 1 and in memory block 1# are rendered conductive, while those of all other memory blocks remain inactive. The unit decoder 302 activates the unit word line UWL2 while keeping other unit word lines inactive. Therefore, only WL12 is activated while all other block word lines remain inactive. The transistor 110 in memory cell MC12 is rendered conductive. Thereby, signal charge in capacitor of memory cell MC12 is read out to block bit line BL12 and to unit bit line UBL2 through the block bit line switch 314. In the mean time, BL12# is also connected to UBL2# through the block bit line switch in memory block 1#, but there is no signal charge read out to UBL2# because WL12# remains inactive. Since the bit lines in the memory block pairs are drawn in mirror symmetry, their parasitic capacitance is matched. The signal charge in memory cell MC12 develops a minute difference of electric potential between UBL2 and UBL2#. Such difference is detected and is amplified by sense amplifier AMP2; the result is sent to high order data bus (not shown), and is used to refresh memory cell MC12. Similarly, the content of memory cell MC11 is read and refreshed by sense amplifier AMP1; the content of memory cell MCi1 is read and refreshed by sense amplifier AMPi.

If we want to read the data from memory cells on WL12# in block 1#, the procedure is identical except that the unit decoder 302 should activate UWL2# instead of UWL2. If we want to read from memory cells in WLij in block i, the unit decoder 302 should turn on UWLj and the block select signal BLKSELi should be activated. The content of memory cell MCi1 is read and refreshed by sense amplifier AMP1; the content of memory cell MCi2 is read and refreshed by sense amplifier AMP2; . . . ; and the content of memory cell MCi1 is read and refreshed by sense amplifier AMPi.

It is still true that one sense amplifier is activated for each activated memory cell; otherwise the data stored in the memory cell will be lost. The differences are that the activated sense amplifiers no long need to be placed right next to the local bit lines connected to the activated memory cell and that the number of activated memory cells is only a small fraction of that of a prior art DRAM. The multiple dimensional bit line structure allows us to place the activated sense amplifier far away from the activated memory cells without introducing excessive parasitic loading to the bit lines. The layout pitches of sense amplifier and peripheral circuits are independent of the size of memory cell. It is therefore possible to design high performance peripheral circuits without increasing the area significantly.

It is to be understood that the present invention describes multiple dimension bit line structure "before" the first level sense amplifiers detect the storage charges in the activated memory cells. Prior art multi-bank DRAM often has multiple dimension data buses "after" the first level sense amplifier already detected the storage charge in activated

memory cells. The prior art multi-bank memories need one first level sense amplifier for every bit line pairs, and they do not solve the tight pitch layout problem.

While specific embodiments of the invention have been illustrated and described herein, it is realized that other modification and changes will occur to those skilled in the art. For example, the above embodiment assumes that bit line pairs are rendered in opposite memory block pairs. It should be obvious to those skilled in the art that this invention also can support the conventional bit line pairing structure in FIG. 1 where bit line pairs are arranged right next to each other. It is also obvious that the above two-dimensional bit line structure can be easily expanded to three-dimensional or multi-dimensional bit line structures. A two dimensional bit line structure is described in FIG. 3a for its simplicity, but the number of levels of bit line structures is not limited by the above example. The optimum levels of bit line structures are determined by details of manufacture technology and by the design specifications.

It also should be obvious that the bit line switches are not required elements; the unit bit lines can be connected directly to block bit lines without bit lines switches. Bit line switches help to reduce the bit line capacitance seen by each sense amplifier, but they are not required for functional reason because the word line switches already can isolate the memory cells in each memory block from memory cells in other memory blocks. While one sense amplifier is placed in each pair of memory block in the above example, there is no such constraint in this invention. We can place more than one sense amplifier per memory block, or place one sense amplifier in the area of many memory blocks. Because of a structure of multiple dimension bit line, the present invention completely removes the layout constraint between memory array and peripheral circuits.

FIG. 3b shows a memory array of the present invention with 3-level bit line connections. For simplicity, only two pairs of bit lines are shown in this figure. The first level of bit lines are made by the first layer metal (M1), the second level is made by the second layer metal (M2), and the third level is made by the third layer metal (M3). Each memory block 350 contains a plurality of side-by-side M1 bit line pairs (BBLi, BBLi#), (BBLj, BBLj#). This memory array contains a plurality of memory columns 360. The M1 bit lines are connected to corresponding M1 bit lines in other memory blocks along the same memory column 360 by M2 bit lines CBLi, CBLi#, CBLj, CBLj#. The bit lines in each column are connected to the bit lines in other columns using metal 3 bit lines M3Li, M3Li#, M3Lj, M3Lj# through bit line switches 362. For each bit line in one memory column 360, we only need one bit line switch 362 and one M3 bit line. A group of sense amplifiers SA1, . . . , Sai, . . . , SAj, are placed at one end of the memory array. Each pair of the above three-dimension bit line networks are connected to one sense amplifier. For example, the (BBLi, CBLi, M3Li), (BBLi#, CBLi#, M3Li#) pair are connected to Sai, and the (BBLj, CBLj, M3Lj), (BBLj#, CBLj#, M3Lj#) pair are connected to SAj. Since each memory block 350 has its own word line switch (not shown in FIG. 3b), no more than one memory block in the network can be activated at any time. It is therefore possible to support a large number of memory cells using a small number of sense amplifiers without violating the requirement that every activated memory cell must have an activated sense amplifier to detect its storage charge.

Although the bit line structure in FIG. 3b is the actual bit line structure used in our product, for simplicity, we will use the simpler two-dimensional bit line structure in FIG. 3a as example in the following discussions.

The difference in layout area and the difference in power consumption between the prior art and this invention are illustrated by the simplified block diagrams in FIGS. 4(a,b). FIG. 4a shows a simplified symbolic graph of one memory bank of conventional DRAM memory array 400 that has N bit line pairs, M word lines, and 8 output (N and M are integers). The sense amplifiers are represented by long rectangles 402 in FIG. 4a. Because one sense amplifier supports each bit line pair, the layout pitch for the sense amplifier is the layout pitch of a bit line pair, so that they must be placed in long narrow rectangular area. The outputs of the sense amplifiers are selected into 8 outputs by the output decoder 404 and multiplexers 406. The layout pitch for the output decoder 404 is also very narrow. The layout pitch for each element of the word line decoder 410 is the pitch of one memory cell Cx. For a memory operation, one word line 412 is activated across the whole memory bank. The number of active memory transistors is N. All N sense amplifiers are activated, and all N bit line pairs in this memory bank are charged or discharged by the sense amplifiers. The activated area covers the whole memory bank as illustrated by the shaded area in FIG. 4a.

FIG. 4b is a simplified symbolic graph of one bank of DRAM memory array of the present invention. For simplicity in comparison, we assume that the memory array in FIG. 4b contains the same number of memory cells and the same number of data outputs as the memory array in FIG. 4a. The memory bank is divided into 4 units 450, and each unit contains 8 pairs of memory blocks 452. We have one amplifier 454 for each pairs of memory blocks. Each unit has one unit word line decoder 456. Detailed structure of the memory unit has been described in FIG. 3a. A unit select decoder 460 generates unit select signals XBLKSEL along word line directions. A block select decoder 462 generates bank level block select signals YBLKSEL. A memory block 452 is activated when both XBLKSEL and YBLKSEL crossing the block are activated. The local block select signals are generated by AND gates in the amplifier 454 area. The outputs of each amplifier is placed on bank level bit lines KBL, KBL# to input/out (IO) units 470 at the edge of the memory. For simplicity, only one pair of bank level bit lines are shown in FIG. 4b. Further details of those peripheral circuits will be discussed in following sections. FIG. 4b shows that the layout pitch for the sense amplifiers 454 is 8 times wider than that in FIG. 4a. The peripheral circuits no longer require tight pitch layout, so that we can design them efficiently for both speed and area consideration. For a memory operation, only one memory block 452 and 8 sense amplifiers 454 in the selected unit 450 are activated. The shaded area in FIG. 4b illustrates the activated area. This active area is obviously much smaller than the active area of a conventional memory bank shown in FIG. 4a. Power consumption of the present invention is therefore much less than that of a prior art memory.

The parasitic bit line parasitic capacitance Cbp of the prior art memory in FIG. 4a is

$$C_{bp} = (M/2) * C_d + M * C_{m1} \quad (1)$$

And, where Cd is the diffusion capacitance for one bit line contact, Cm1 is the metal 1 capacitance of the bit line for each unit cell, and M is the number of memory cells along one bit line. We assume that two memory cells share each contact so that the total number of contacts is M/2.

The parasitic bit line capacitance Cb of the memory in FIG. 4b is

$$C_b = (M/16) * C_d + (M/8) * C_{m1} + (8 * C_d + N * C_{m2}) \quad (2)$$

where Cm2 is the metal 2 bit line capacitance for each memory pitch along the unit bit line direction. The first

two terms  $(M/16) \cdot C_d + (M/8) \cdot C_{m1}$  are the capacitance for a local bit line that is  $1/8$  of the length of the bit line in FIG. 4a. The last two terms  $(8 \cdot C_d + N \cdot C_{m2})$  are the parasitic capacitance of the unit bit line that has 8 contacts to the bit line switches and a metal 2 bit line. The contact capacitance  $C_d$  is much larger than the metal 2 capacitance  $C_{m2}$  is usually smaller than the metal 1 capacitance  $C_{m1}$ . Therefore, Eqs. (1,2) show that the bit line parasitic capacitance seen by one sense amplifier of the present invention,  $C_b$ , is significantly smaller than  $C_{bp}$ . Smaller bit line capacitance implies faster speed, lower power, and better reliability. There is no need to use complex technology to build the memory cells. It is also possible to increase the size of each memory block to connect more memory cells to each sense amplifier in order to reduce the total area.

The total areas occupied by memory cells are identical between the two memory arrays in FIG. 4a and FIG. 4b. Therefore, the difference in area is completely determined by the layout of peripheral circuits. The available layout pitch for sense amplifiers and for output decoders for the memory in FIG. 4b is 8 times larger than that of the memory in FIG. 4a. It should be obvious to those skilled in the art that a memory of the present invention is smaller than a prior art memory along the dimension vertical to the word line direction due to wider layout pitch. Along the dimension in parallel to word lines, the present invention still needs a decoder 460 of the same layout pitch. In addition, this invention needs to have one set of word line switches 462 for each memory block 452. The additional area occupied by the word line switches 462 does not increase the layout area significantly because we can use smaller high level decoders due to reduction in loading.

The sense amplifier used in the present invention is substantially the same as typical sense amplifiers used in the prior art. FIG. 5 shows schematic diagram of the amplifier in FIG. 3a. When the sense amplifier enable signal SAEN is activated, transistors MP11, MP12, MN11, and MN12 form a small signal sensing circuit that can detect minute potential difference on the unit bit line pairs UBL and UBL#. The transfer gate transistor MN14 transfers the signal between the unit level bit line UBL and the bank level bit line KBL when the bank level word line KWL is active. The transfer gate transistor MN13 transfers the signal between the unit level bit line UBL# and the bank level bit line KBL# when the bank level word line KWL is active. MN17 is used to equalize the voltages on UBL and UBL# when the sense amplifier is not active. The operation principles of the above sense amplifiers are well known to the art of memory design so we do not describe them in further details.

FIG. 6 is a block diagram of the IO unit 470 in FIG. 4b. The bank level bit line pair KBL and KBL# are connected to a bank level sense amplifier 650 through a bank level bit line switch 651. This sense amplifier 650 is identical to the sense amplifier in FIG. 5; its enable signal is KSAEN. The KBL switch 651 is rendered conductive when its enable signal MREAD is active, and it isolates the bit lines from the sense amplifier when MREAD is not active. This bit line switch 651 is used to improve the speed of the sense amplifier as well known to the art of memory design. The output of the sense amplifier, SOUT, is connected to an Error-Correction-Code (ECC) circuit 652. The ECC circuit is well known to the art, so we do not discuss it in further details. The output of the ECC circuit, EOUT, is connected to the input of an output driver 665. The output driver 665 drives the data to external pad when it is enabled by the

signal READOUT. For a write operation, we place the data on the pad into a storage register 662. The output of the storage register, UDATA, is connected to a memory write driver 664. The memory write driver 664 is controlled by the UPDATE signal to drive data on KBL and KBL# during a memory update operation.

FIGS. 7(a-c) show the waveforms of critical signals for the memory described in previous sections.

FIG. 7a shows the timing of critical signals during a memory operation to read data from memory cells (called a "read cycle"). First, the block select signal BLKSEL is activated at time T1. BLKSEL is active when both XBLKSEL and YBLKSEL are active. Whenever BLKSEL is active, the precharge circuit of the selected memory block is turned off, so does the precharge circuit of all the sense amplifiers of the selected memory unit. The precharge signal and bank level block select signals XBLKSEL, YBLKSEL are not shown in waveforms because the information is redundant with respect to BLKSEL signal. After BLKSEL is active, block word line WL is active at time T2. Once WL is active, a minute potential difference starts to develop in block bit line pair BL, BL# as well as unit bit line pair UBL, UBL#. After enough potential difference has developed on the unit bit line pairs, the sense amplifiers of the selected memory unit are activated by bring SAVCC to VCC, and SAVSS to VSS. The unit sense amplifier starts to magnify the bit line potential once it is activated at T3. The bank level word line KWL is then activated at T4; the potential differences in UBL and UBL# are transferred to bank bit line pairs KBL and KBL# once KWL is activated. Between time T4 and T5, the voltages of UBL and UBL# are first drawn toward PCGV due to charge sharing effect between bank bit lines and unit bit lines; the unit sense amplifier eventually will overcome the charge sharing effect and magnify their potential difference. At time T5, the bank-word-line KWL is off, and the pulling of KSAVCC to VCC and KSAVSS to VSS activates the bank level sense amplifier 750. The bank level sense amplifier 750 will magnify the potential difference on KBL and KBL# to full power supply voltages. In the mean time, the unit level sense amplifier will also pull UBL and UBL# to full power supply voltage. Because we are relying on the unit level sense amplifier to refresh the selected memory cells, we need to provide a timing margin to make sure the signal charges in those memory cells are fully restored before we can turn off the word line WL at T6. After the word line is off, sense amplifiers are deactivated at T7, then the block select signal BLKSEL is deactivated at T8. Once BLKSEL is off, the memory is set into precharge state, and all bit line voltages return to PCGV. A memory of this invention has much shorter precharge time than prior art memories due to much lower loading on each level of its bit lines. At time T9, all signals are fully restored to their precharge states, and the memory is ready for next memory operation.

FIG. 7b shows the timing of critical signals for a memory operation to refresh the data of memory cells (called a "refresh cycle"). A refresh cycle is very similar to a read cycle except that we do not need to bring the data to bank level. All these bank level signals, KWL, KSAVCC, KSAVSS, KBL, and KBL# remain inactive throughout a refresh cycle. At time T11, the block select signal BLKSEL is active, then the word line WL is activated at time T12. Potential differences start to develop in block level and unit level bit lines BL, BL#, UBL, and UBL#. At time T13 the sense amplifier is activated. The sense amplifier quickly magnify and drive the bit lines to full power supply voltages. When the charges in selected memory cells are fully



restored, we can turn off the word line WL at T14, then turn off block select signal BLKSEL at T15. At time T16, all the signals are restored into precharge states, and the memory is ready for next operation. Comparing FIG. 7b with FIG. 7a, it is obvious that the time need for a fresh cycle is shorter than the time for a read cycle because we do not need to drive KBL and KBL#.

FIG. 7c shows the timing of critical signals during a memory operation to write new data into memory cells (called a "write cycle"). At time T21, the block-select-signal BLKSEL and bank level word line KWL are activated. In the mean time, the new data is written into the bank level bit lines KBL and KBL#, then propagate into lower level bit lines UBL, UBL#, BL, and BL#. The memory write driver 764 has strong driving capability so that bit lines can be driven to desired values quickly. At time T22, the unit level sense amplifier is activated to assist the write operation. Once the charges in the memory cells are fully updated, the word lines WL and KWL are turned off at T23. Then, the block select signal BLKSEL are off at T24. At T25 the memory is fully restored to precharge state ready for next memory operation. Comparing FIG. 7c with FIG. 7a, it is obvious that the time needed to execute a write cycle is much shorter than the time needed to execute a read cycle because of the strong driving capability of the memory write driver 764.

As illustrated by FIG. 7a, the reason why read operation is slower than write or refresh operations is because the read operation cannot be finished until the unit level sense amplifiers fully restore the signal charges in the selected memory cells. From the point of view of an external user, the additional time required to refresh the memory does not influence the total latency for a memory read operation because the process to deliver data from bank level circuit to external pad is executed in parallel. The refresh time is therefore "hidden" from external users. The only time an external user can feel the effect of this additional refresh time is when a refresh cycle is scheduled at the same time as a read cycle is requested. The memory can not execute a refresh cycle in parallel to a read cycle at a different address, so one of the requests must wait. External control logic is therefore necessary to handle this resource conflict condition. For a memory with ECC support, data write operations always need to start with memory read operations, so the above problems also apply to memory write operations. In order to fully compatible with an SRAM, we must make internal memory refresh cycles completely invisible to external users. This is achieved by simple changes in IO circuit shown in FIG. 8, and change in timing control shown in FIG. 9.

The IO circuit in FIG. 8 is almost identical to the IO circuit in FIG. 6 except that it has two additional multiplexers 854, 860. The output of the ECC circuit, EOUT, is connected to the input of a bypass multiplexer 854. During a read cycle, the bypass multiplexer 854 selects the output from the storage register 662 if the reading memory address matches the address of the data stored in the storage register 662. Otherwise, the bypass multiplexer 854 selects the output of the ECC circuit, and sends the memory output to the output driver 665. The storage multiplexer 860 selects the input from external pad during a write operation, and it selects the data from memory read out during a read operation. This architecture allows us to "hide" a refresh cycle in parallel with a normal memory operation. It also improves the speed of normal read operations. Using the circuit in FIG. 8, the most updated data of previous memory operation are always stored into the storage register 662. To execute a

new memory operation, we always check if the data are stored in the storage register before reading data from the memory array. If the wanted data is already stored in the storage register, no memory operation will be executed, and the data is read from the storage register directly. When a new set of data is read from the memory array, an update cycle is always executed before the end of a new memory operation to write the data currently in the storage buffer back into the memory array. Since we always store every memory read results into the storage registers, there is no need to refresh the selected memory cells immediately. With this configuration, we can terminate the read operation before the unit level sense amplifier can fully refresh the activated memory cells. Therefore, the unit level circuits are available for a refresh cycle at the same time when the memory is propagating the read data to the external pads. This architecture removes the conflict between refresh cycle and normal memory operations. The operation principle of this scheme is further illustrated by the waveforms in FIG. 9.

FIG. 9 shows the worst case situation when a memory operation overlaps with a refresh operation (to a different address or to the same memory block), and when there is a need to update data from the storage buffer at the same time. Under this worst case condition, the refresh cycle and the memory update cycle must be "hidden" in the memory read operation in order to avoid complexity in system support. On the other word, we must execute the refresh and update cycles in parallel without influencing the timing observable by an external user.

At time Tr1 in FIG. 9, the block select signal BLKSEL is activated for a read operation. At time Tr2, the word line WL is activated, then the unit sense amplifier is activated at Tr3. The unit level word line KWL is activated at Tr4, and the unit level sense amplifier is activated at time Tr5. Until time Tr5, the memory operations and waveforms are identical to those shown in the read cycle in FIG. 8a. The operation is different starting at Tr5; we are allowed to turn off the block select signal BLKSEL, the word lines WL, KWL, and the unit level sense amplifier simultaneously at Tr5 without waiting for full amplification of the memory data. The memory block quickly recovers to precharge state ready for next operation at time Tf1. During this time period, the unit level sense amplifier does not have enough time to fully amplify the signals in the lower level bit lines BL, BL#, UBL, and UBL#. Those activated memory cells no longer stores the original data. That is perfectly all right because the correct data will be stored in the storage register 662 in the following procedures. At time Tf1, the data are sensed by the bank level sense amplifier; the correct data will be remembered in the storage register 662 and updated into those selected memory in the next memory operation. Therefore, the data are not lost even when the storage charge in the memory cells are neutralized at this time. At the same time when we are waiting for the bank level circuits to propagate the new read data to external circuits, the unit level and block level memory circuits are available for a refresh operation. This hidden refresh cycle can happen at any memory address. The worst case timing happen when the refresh cycle happens at the same block that we just read. FIG. 9 shows the timing of the worst case condition. At time Tf1, BLKSEL is activated for the refresh cycle. A refresh cycle with identical waveforms as the waveforms in FIG. 8b is executed from time Tf1 to time Tf5. At time Tw1, the memory unit is ready for new operation, and the bank level read operation is completed. At this time, the IO unit 720 is executing ECC correction and the data is propagating to the

pads. In the mean time, the bank level resources are available, so we take this chance to update the old data in the storage register 762 back into the memory array by executing a write cycle. The waveforms in FIG. 9 from time Tw1 to Tw5 are identical to the waveforms in FIG. 7c. At the end of the memory operation, the latest data just read from the memory are stored into the storage register 662, the previous data are updated into the memory array, we fulfilled a refresh request, and the external memory operation request is completed.

It is still true that we need to record the data stored in every activated memory cell. Otherwise the data will be lost. The difference between the above memory access procedures and conventional DRAM memory accesses is that the data is temporarily stored in the storage registers so that we do not need to refresh the activated memory cells immediately. This architecture delays data update until next memory process using available bandwidth, so that refresh cycles and update cycles can be hidden to improve system performance.

The above architecture is different from a hybrid memory because (1) this invention simplifies the timing control of DRAM read cycle while the SRAM of the hybrid memory does not simplify the DRAM operation, (2) the system control and device performance of the present invention is the same no matter the memory operation hits the storage register or not, while the performance and control of a cache memory is significantly different when the memory operation miss the cache array, (3) a hybrid memory has better performance when the size of the SRAM cache is larger due to higher hit rate, while the performance of the present invention is independent of hit rate, and (4) the storage register does not introduce significant area penalty while the on-chip SRAM of hybrid memory occupies a significant layout area. The structure and the operation principles of the memory architecture described in the above sections are therefore completely different from the structures of hybrid memories.

As apparent from the foregoing, the following advantages may be obtained according to this invention.

- (1) The tight pitch layout problem is solved completely. Since many bit line pairs share the same sense amplifier, the available layout pitch for each peripheral circuit is many times of the memory cell pitch. Therefore, sense amplifiers and peripheral circuits of high sensitivity with electrical symmetry and high layout efficiency can be realized.
- (2) The bit line loading seen by the sense amplifier is reduced dramatically. It is therefore possible to improve the performance significantly.
- (3) It is also possible to attach a large number of memory cells to each sense amplifier to reduce total device area.
- (4) The novel design in decoder reduces decoder size significantly without sacrificing driving capability. The loading on each unit word line is also reduced significantly. This decoder design reduces layout area and improves device performance.
- (5) Changes in memory access procedures allow us to delay the refresh operation until next memory operation. Internal refresh operations are therefore invisible for external users.
- (6) The only devices activated in each memory operation are those devices must be activated. There is little waste in power. The present invention consumes much less power than prior art memories.

A memory device of the present invention is under production. Using 0.6 micron technology to build a memory

array containing one million memory cells, we are able to achieve 4 ns access time, which is more than 10 times faster than existing memories devices of the same storage capacity.

FIG. 10 shows an example of a typical prior art decoder. Each branch of the decoder contains one AND gate 1101 that controls one of the outputs of the decoder O3-0. Two sets of mutually exclusive input select signals (G0, G0NN) and (G1, G1NN) are connected to the inputs of those AND gates as show in FIG. 10, so that no more than one output O3-0 of the decoder can be activated at any time.

FIG. 11(a) is the schematic diagram of a single-transistor decoder that uses only one n-channel transistor M3 to M0 for each branch of the decoder. The source of each transistor M3 to M0 is connected to one word line WL3 to WL0 of the memory array. A set of mutually exclusive drain select signals DSEL1, DSEL0 are connected to the drains of those transistors M3 to M0, and a set of mutually exclusive gate select signals GSEL1 and GSEL0 are connected to the gates of those transistors M3 to M0, as shown in FIG. 11(a). In this configuration, WL3 is activated only when both DSEL1 and GSEL1 are activated, WL2 is activated only when both DSEL1 and GSEL0 are activated, WL1 is activated only when both DSEL0 and GSEL1 are activated, and WL0 is activated only when both DSEL0 and GSEL0 are activated. Therefore, the circuit in FIG. 11(a) fulfills the necessary function of a memory word line decoder. A typical CMOS AND gate contains 3 p-channel transistors and 3 n-channel transistors. The decoder in FIG. 12(a) uses only one transistor for each output of the decoder. It is apparent that the decoder in FIG. 11(a) is by far smaller than the one in FIG. 10. However, the single-transistor decoder in FIG. 11(a) requires special timing controls as illustrated in the following example.

FIG. 11(b) illustrates the timing of input signals to activate one of the word line WL0. Before time T0, there are no decoding activities. All gate select signals GSEL1, GSEL0 must stay at power supply voltage Vcc, and all drain select signals DSEL1, DSEL0 must stay at ground voltage Vss. Otherwise one of the word line maybe activated accidentally by noise or leakage. To activate one word line WL0, we must deactivate all gate select signals GSEL1, GSEL0 at time T0, then activate one of the gate select signal GSEL0 and one of the drain select signal DSEL0 at T1. In order to deactivate the decoder, DSEL0 must be deactivated at time T2 before all gate select signals GSEL1 and GSEL0 are activated again at T3. The above control sequence is necessary to prevent accidental activation of word lines that are not selected. The above timing control sequence is complex because all inputs are involved when we only want to active one word line. The above decoders are simplified examples of 4 output decoders. A realistic memory decoder will need to control thousands of word lines. The power consumed by such complex control sequences can be significant for a realistic memory decoder. Another problem for the decoder in FIG. 11(a) is also illustrated in FIG. 11(b). Due to body effect of n-channel transistor M0, the voltage of the activated word line WL0 is lower than the power supply voltage Vcc by an amount Vbd as shown in FIG. 11(b). This voltage drop can be a big problem for a DRAM decoder because it will reduce the signal charge stored in DRAM memory cells.

FIG. 12(a) is a schematic diagram of a decoder of the present invention. The only differences between the decoders in FIGS. 11(a), 12(a) is that depletion mode transistors D3 to D0, instead of enhanced mode transistors M3 to M0, are used by the decoder shown in FIG. 12(a). The threshold voltage of those depletion mode transistors D3 to D0 is controlled to be around -0.2 volts (or roughly 1/3 of the

US 6,687,148 B2

17

threshold voltage of a typical enhance mode transistor) below power supply voltage Vss.

FIG. 12(b) illustrates the timing of input signals to select one word line WL0 of the depletion-mode single transistor decoder in FIG. 12(a). Before time T0, all the gate select 5 singels GSEL1, GSEL0, and all the drain select signals DSEL1, DSEL0 are at ground voltage Vss. Unlike the enhance mode single transistor decoder in FIG. 11(a), it is all right to set the gate control signals GSEL1, GSEL0 at Vss when the decoder is idle. The word lines WL3-WL0 won't be activated by noise or small leakage because the depletion-mode transistors D3 to D0 are partially on when its gate voltage is at Vss. To activate one word line WL0, we no longer need to deactivate all gate select signals. We only need to activate one gate select signal GSEL0 and one drain 15 select signal DSEL0 as shown in FIG. 12(b). To deactivate the decoder, we can simply deactivate GSEL0 and DSEL0 as shown in FIG. 12(b). This control sequence is apparently much simpler than the control sequence shown in FIG. 11(b). There is also no voltage drop cause by body effect on the selected word line because the threshold voltage of the activated transistor M0 is below zero. The depletion mode single transistor decoder in FIG. 12(a) is equally small in area as the enhance mode single transistor decoder in FIG. 11(a), but it will consume much less power. The only 25 problem is that some of those word lines are partially activated when they have deactivated gate select signal but activated drain select signal as illustrated by WL1 in FIG. 12(b). This partial activation of word lines is not a functional problem when the voltage Vpt is less than the threshold voltage of selection gates in the memory cells, but it may introduce a potential charge retention problem due to sub-threshold leakage current. One solution for this problem is to introduce a small negative voltage on all deactivated gate 35 select signals at time T0 as shown in FIG. 12(c). This small negative voltage Vnt on the drain select signal assures the depletion gate transistor D1 remains uncondutive so that the word line WL1 won't be partially activated.

While specific embodiments of single transistor decoders have been illustrated and described herein, it is realized that other modifications and changes will occur to those skilled in the art. For example, p-channel transistors or depletion mode p-channel transistors can replace the n-channel transistors in the above examples.

As apparent from the foregoing, single-transistor-decoders of the present invention occupies much small area than the prior art CMOS-decoders. It is therefore possible to divide a large memory array into small block—each block isolated by its own decoder—without increasing the total area significantly. When the memory array is divided into small blocks, we no longer need to have large storage capacitor as prior art DRAM cells have. It is therefore possible to build DRAM memory cells using standard logic technology.

One example of DRAM memory cell built by logic technology is shown in FIG. 13. This memory cell 1400 contains one select transistor 1402 and one storage transistor 1404. The gate of the storage transistor 1404 is biased to full power supply voltage Vcc so that it behaves as a capacitor. The drain of the storage transistor 1404 is connected to the source of the select transistor 1402. The gate of the select transistor 1402 is connected to word line WL, and the drain of the select transistor is connected to bit line BL. Using this memory cell 1400 and a memory architecture disclosed in this invention and in our previous patent application, commercial memory products were manufactured successfully. The major advantage of the logic memory cell 1400 is that

18

it can be manufactured using standard logic technology. The resulting memory product achieved unprecedented high performance. The area of the logic memory cell 1400 is larger than prior art DRAM cells because two transistors, instead of one transistor and one capacitor, are used to build one memory cell. It is therefore desirable to be able to build single transistor memory cell from a manufacture technology as similar to logic technology as possible.

Therefore, according to FIGS. 3a to 4b, and FIGS. 12(a) to 13, a semiconductor memory device 300 is disclosed which is provided for operation with a plurality of cell-refreshing sense-amplifiers (SAs). The memory device 300 includes a memory cell array having a plurality of first-direction first-level bit lines, e.g., bit line BLni in block n for bit-i, along a first bit-line direction, disposed in a parallel manner along a first direction, e.g., a horizontal direction. The memory cell array further includes a plurality of word lines WL intersected with the first-direction first-level bit lines. The memory cell array further includes a plurality of memory cells. Each of these plurality of memory cells being coupled between one of the first-direction first level bit lines, i.e., bit line BLni in block n for bit-i, along a first bit-line direction and one of the word lines for storing data therein. The memory device further includes a plurality of different-direction first level bit lines, e.g., multiple-block or the unit bit-line-i such as UBLi, BBLi, CBLi, etc. (referring to FIG. 3b), where i=1, 2, 3, . . . I, disposed along a plurality of different directions, e.g., along a vertical direction, with at least one of the different directions being different from the first direction, wherein each of the first direction first level bit lines connected to one of the cell-refreshing sense amplifiers (SAs) directly or via the different-direction first level bit-lines. In a specific preferred embodiment, one of the different directions, e.g., a vertical direction, for arranging the different-direction first level bit lines, e.g., the multiple-block bit-line-i UBLi, BBLi, CBLi, etc. (referring to FIG. 3b). Where i=1, 2, 3, . . . I, being perpendicular to the first direction, e.g., a horizontal direction for arranging the first-direction first level bit lines. In the preferred embodiment as shown in FIG. 4b, the memory device 300 further includes bit line switches connected between the first level bit lines, which are arranged in different directions. The semiconductor memory device further includes a decoder 302 for generating an activating signal for activating one of the word lines WL. The decoder 302 further includes a plurality of drain select lines, e.g., DSEL0 AND DSEL1, etc., each being provided for receiving one of a plurality of mutual exclusively drain select signals. The decoder 302 further includes a plurality of gate select lines, e.g., GSEL0, GSEL1, etc., each being provided for receiving one of a plurality of mutual exclusively gate select signals. The decoder 302 further includes a plurality of transistors, e.g., D0, D1, or M0, M1, etc. Each transistor includes a drain which being connected correspondingly to one of the plurality of drain select input lines, e.g., DSEL0, DSEL1, etc., for receiving one of the mutually exclusive drain select signals therefrom. Each of the transistors further includes a gate which being connected correspondingly to one of the plurality of gate select input lines GSEL0, GSEL1, etc., for receiving one of the mutually exclusive gate select signals therefrom. Each of the plurality of transistors further includes a source, which is connected to an output signal line for providing the activating signal to one of the word lines WL which being contingent upon the mutually exclusive drain select signals DSEL0, DSEL1, etc. And, the mutually exclusive gate select signals GSEL0, GSEL1, etc. In a preferred embodiment, each of the transistors is an enhanced



mode transistor, and in another preferred embodiment, each of the transistors is a depletion mode transistor.

Furthermore, according to FIGS. 3a to 4b, and FIGS. 12(a) to 13 a method for configuring a semiconductor memory device for operation with a plurality of cell-refreshing sense-amplifiers (SAs) is also disclosed. The method includes the steps of (a) arranging a plurality of first-direction first-level bit lines in a parallel manner along a first direction; (b) arranging a plurality of word lines for intersecting with the first-direction first-level bit lines; (c) coupling a memory cell between each of the first-direction first level bit lines and one of the word lines for storing data therein; (d) arranging a plurality of different-direction first level bit lines along a plurality of different directions with at least one of the different directions being different from the first direction; (e) connecting each of the first direction first level bit lines to one of the cell-refreshing sense amplifiers (SAs) directly or via the different-direction first level bit-lines; (f) connecting each of the word lines WL to a decoder 302 for receiving an activating signal therefrom for activating one of the word lines WL; (g) forming the decoder with a plurality of transistors each includes a drain, a gate and a source therein; (h) connecting a drain select line to each of the drain of each of the transistors and connecting a gate select line to each of the gate of each of the transistors; (i) applying each of the drain select lines to receive one of a plurality of mutually exclusive drain select signals and each of the gate select lines to receive one of a plurality of mutually exclusive gate select signals; and (j) applying each of the plurality of transistors to generate an output signal from each of the source which being contingent upon the mutually exclusive drain select signals and the mutually exclusive gate select signals for providing the activating signal to each of the word lines.

According to FIG. 13, this invention further discloses a dynamic random access memory (DRAM) cell which is coupled to a word-line and a bit-line. The DRAM memory cell includes a select transistor 1402 includes a drain connected to the bit line BL and a gate connected to the word line WL. The cell further includes a storage transistor 1404 includes a drain connected to the source of the select transistor 1402 and a gate connected to a power supply voltage Vcc whereby the storage transistor 1404 is implemented as a capacitor for storing a binary bit therein. In summary, the present invention further discloses a memory cell coupled to a word-line and a bit-line. The memory cell includes a storage transistor connected to the word line and bit line via a select means provided for selectively activating the memory cell. And, the storage transistor further includes a gate, which is biased to a power supply voltage to function, as a capacitor for storing a binary bit therein.

FIGS. 14(a-f) and FIGS. 15(a-c) illustrates a procedure to manufacture high density memory using a manufacture technology very similar to standard logic technology. The first step is to define active area 1502, and grow isolation field oxide 1504 to separate those active area as show in the cross section diagram in FIG. 14(a) and the top view in FIG. 15(a). This step is identical to any standard IC technology. The next step is to apply a mask 1506 to define the location of trench capacitors as illustrated by FIG. 14(b). Selective plasma etching is used to dig a trench 1510 at the opening defined by the field oxide 1504 and the trench mask 1506 as illustrated in the cross-section diagram in FIG. 14(c) and the top view in FIG. 15(b). This is a self-aligned process because three edges of the trench 1510 are defined by field oxide. The trench mask 1506 only needs to define one edge of the trench. After the above processing steps, all the

following processing procedures are conventional manufacture processes of standard logic technology. First, a layer of thin insulator 1511 is grown at the surface of the active area 1502, including the surfaces of the trench 1510 as shown in FIG. 14(d). The next step is to deposit poly silicon 1512 to fill the trench 1510 and cover the whole silicon as illustrated in FIG. 14(e). A poly mask 1520 is then used for poly silicon etching process to define transistor gates 1522 and the electrode 1524 of the trench capacitor as illustrated in FIG. 14(f). FIG. 15(c) shows the top view and FIG. 14(g) shows the cross-sectional view of the resulting memory cell structure. The trench capacitors 1510 are filled with poly silicon. One electrode 1602 of all those trench capacitors 1510 are connected together through poly silicon to power supply voltage Vcc. The other electrodes of the trench capacitors are connected to the sources of select transistors 1604. The poly silicon word lines 1606 define the gates of the select transistors, and the drains of the select transistors are connected to metal bit lines through diffusion contacts 1608.

As apparent from the foregoing, following advantages are obtained according to this invention.

- (1) All the procedures used to build the DRAM cell are existing procedures of standard logic technology, except one masking step and one plasma-etching step. Comparing with current art embedded memory technologies, the present invention simplifies the manufacture technology by more than 30%.
- (2) The procedure to define the dimension of trench capacitor is a self-aligned procedure; three edges of the trench capacitor are defined by field oxide; only one edge is defined by mask. This self-aligned procedure allows us to minimize the area of the memory cell.

Another procedure has also been developed to build self-aligned trench capacitors using logic technology. The first step is to build CMOS transistors following standard logic technology as illustrated in the cross-section diagram in FIG. 16(a). At this time, the MOS transistor has been fully processed. The poly silicon gate 1702 is already covered by oxide for protection. A trench mask 1706 is then deposited. This trench mask 1706 is used to protect area where we do not want to dig trench capacitor; it is not needed to define the dimension of the trench capacitor because all four edges of the area are already defined. Three edges are defined by the field oxide 1710 in the same way as the previous procedure, and the forth edge is define by the oxide 1704 on the transistor gate. This is therefore a complete self-aligned procedure. The following selective plasma etching procedure is therefore able to utilize optimum area for the trench capacitor as illustrated in FIG. 16(b). Thin insulation layer is grown on the surfaces of the trench 1712 before the whole area is covered by second layer poly silicon 1714 as shown in FIG. 16(c). Photo resist 1716 that is defined by the same mask as the one used in FIG. 16(a) defines the dimension of the second layer poly silicon 1716 (the polarity of the photo resist used in FIG. 16(a) is opposite to that used in FIG. 16(c)). The second layer poly silicon 1716 is then etched to form the electrodes 1720 of those trench capacitors 1722. FIG. 17 shows the top view of the DRAM cells manufactured by the above procedures. The word lines 1802 are defined by the first layer poly silicon. Second layer poly silicon are used to fill the trench capacitors 1722, and to connect one electrode 1720 of all those trench capacitors to Vcc.

The above procedure is more complex than the procedure illustrated in FIGS. 14(a-g). It has the advantage that the trench capacitors are fully self-aligned for all 4 edges of their opening. Utilization of the silicon area is therefore fully

21

optimized. While specific embodiments of the invention have been illustrated and described herein, it is realized that other modification and changes will occur to those skilled in the art. For example, the insulation-layer in the trench capacitors maybe grown in a different processing step instead of during the process of forming the gate oxide. The exact sequence of the processing steps also can be varied to achieve similar simplification.

The top electrode (1602) of the trench capacitor (1510) of the memory cells shown in FIG. (14) must be connected to a voltage at least one threshold voltage ( $V_t$ ) higher than the voltage of the bottom electrode to make the area under the insulator layer (1511) conductive. Similarly, the top electrode (1702) of the trench capacitor of the memory cells shown in FIG. (16) also must be connected to a voltage at least one  $V_t$  higher than the voltage of the bottom electrode. Typically, those top electrodes (1602,1702) are connected to power supply voltage  $V_{cc}$ . This constraint can be removed if a diffusion layer (1805) is deposited around the trench capacitor (1802) as illustrated by the cross-section diagram in FIG. 18(a). This diffusion layer (1805), the drain of the word line transistor (1606), and the top electrode (1602) are all doped with the same type of doping. Therefore, the bottom electrode of the trench capacitor (1801) is always conductive, which removes the constraint on the electrode voltages. The cross-section diagram in FIG. 18(b) illustrates another variation in device structure. In this structure, a transistor (1811) instead of field oxide separates two nearby trench capacitors (1821, 1823). The gate (1813) of this isolation transistor (1811) is connected to ground voltage  $V_{ss}$  to separate nearby trench capacitors (1821, 1823). Transistors (1811, 1815) therefore define two edges of the areas of the trench capacitors (1821, 1823) instead of field oxide, which usually helps to reduce the size of memory cells.

In the above examples, the geometry of memory cell structures is drawn in 90-degree angles for simplicity. In reality, memory cells are often drawn in multiple angles as illustrated by the top view memory cell structures in FIG. 19. The trench capacitors (1901) are placed in 45 degree to the contacts (1903). The word line (1907) and the diffusion area (1905) are also placed in 45-degree angles. Since the area of the trench capacitors (1901) are defined by field oxide and transistor edges, its shape is therefore not necessary rectangular as shown by the example in FIG. 19.

The word line transistor (1402) in the memory cell of the present invention has the same properties and it is manufactured in the same time as the transistors used for peripheral circuits and logic circuits. The word line transistors of prior art DRAM technologies are always different from logic transistors. In order to tolerate higher word line voltage introduced by the word line boosting circuits, the gate oxide thickness ( $T_{ox}$ ) of a prior art word line transistor is thicker than that of a logic transistor. In order to reduce leakage current, the threshold voltage ( $V_t$ ) of a prior art word-line-transistor is higher. Table 1 lists transistor properties for a typical 0.35  $\mu m$  DRAM technology. The word line transistor and the logic transistor in this example is manufactured by the same procedures except that one masking step is added to increase  $V_t$  of the word line transistor. The word line transistor has higher  $V_t$  (1.1 volts for the example in Table 1) so that it can be drawn to a smaller minimum channel length ( $L_{min}$ ), which is 0.35  $\mu m$  in this case, without leakage problems. The logic transistor has lower  $V_t$  (0.7 volts for this example), but its  $L_{min}$  is larger. On the other word, the logic transistors of a typical DRAM technology is equivalent to the logic transistors of 0.5  $\mu m$  technology

22

instead of 0.35  $\mu m$  technology. On the other word, the performance of logic transistors of DRAM technology is one generation behind the transistors of typical logic technology.

One method to have both high performance logic transistors and low leakage DRAM transistors on the same chip is to make different kinds of transistors using complex manufacture procedures. Table 2 shows the transistor properties for one example of such complex embedded memory technology. This technology has word line transistor with high  $V_t$  and thick oxide, high voltage transistors with thick oxide and long channel length, and logic transistors with low  $V_t$  and thin oxide. The manufacture procedures for such technology are very complex. The manufacture cost is very high.

TABLE 1

Transistor properties for word line transistors and logic transistors of prior art DRAM technology.			
	$T_{ox}$	$V_t$ (volts)	$L_{min}$ (micrometers)
Word line transistor	100	1.1	0.35
Logic transistor	100	0.7	0.5

TABLE 2

Transistor properties for word line transistors and logic transistors of prior art embedded DRAM technology.			
	$T_{ox}$	$V_t$ (volts)	$L_{min}$ (micrometers)
Word line transistor	100	1.1	0.35
High Voltage transistor	100	0.7	0.5
Logic transistor	70	0.7	0.35

TABLE 1

Transistor properties for word line transistors and logic transistors of prior art DRAM technology.			
	$T_{ox}$	$V_t$ (volts)	$L_{min}$ (micrometers)
Word line transistor	100	0.7 (1.1)	0.35
Logic transistor	100	0.7	0.35

A DRAM (dynamic random access memory) cell array supported on a substrate is therefore disclosed in this invention. The DRAM cell array includes a plurality of memory cells each having a select-transistor wherein each of the select-transistor having a select-transistor-gate. The DRAM cell array further includes a peripheral logic-circuit having logic-transistors wherein each of the logic-transistors having a logic-transistor-gate. The select-transistor-gate and the logic-circuit-gate have substantially a same thickness. And, the select-transistor for each of the memory cells having a select-transistor threshold voltage and each of the logic-transistors of the peripheral logic-circuit having a logic-transistor threshold voltage wherein the select-transistor threshold voltage is substantially the same as the logic-transistor threshold voltage. In a preferred embodiment, each of the memory cells further having a trench capacitor. In another preferred embodiment, the DRAM cell array further includes an active area isolated and defined by edges

US 6,687,148 B2

23

of a field oxide layer disposed on the substrate wherein each of the trench capacitors disposed in the active area and in self-alignment with the edges of the field oxide layer. In another preferred embodiment, the DRAM cell array further includes an active area isolated and defined by edges of a field oxide layer disposed on the substrate. Each of the trench capacitors is disposed in the active area and in self-alignment with the edges of the field oxide layer and edges of the select-transistor gate. In another preferred embodiment, the DRAM cell array further includes an error code checking (ECC) and correction means connected to the memory cell array for checking and correcting substantially all memory read errors within a threshold error-detection-and-correction time.

According to above description, this invention discloses a method for manufacturing a DRAM (dynamic random access memory) cell array each having a select-transistor and peripheral logic circuit having logic-transistors supported on a substrate. The method includes the steps of (a) applying a gate-formation process for simultaneously forming a select-transistor-gate for the select-transistor and a logic-circuit-gate for each of the logic-transistors for the peripheral logic-circuit wherein the select-transistor-gate and the logic-circuit-gate having substantially a same thickness; and (b) applying substantially same implant processes in forming the select-transistor and the logic-transistors wherein the select-transistor and the logic transistors having substantially a same threshold voltage. In a preferred embodiment, the method further includes a step of (c) applying a capacitive-transistor trench mask for etching a plurality of trench capacitors for the memory cell array. In a preferred embodiment, the step of applying a capacitive-transistor trench mask is a step of applying a capacitive-transistor trench mask in an active area isolated by a field oxide. The capacitive-transistor trench mask cooperates with the field oxide for etching the trench in self-alignment in the active area with etching edges defined by the field oxide. In another preferred embodiment, the step of applying a capacitive-transistor trench mask in corporation with the field oxide is a step of applying a capacitive-transistor trench mask in an active area isolated by the field oxide as an enclosed area. The capacitive-transistor trench mask is employed to define a single edge of the trench capacitor while remaining edges of the trench capacitor are in self-alignment with the field oxide wherein the etching edges for the remaining edges are inherently defined in the active area by the field oxide. In another preferred embodiment, the step of applying a capacitive-transistor trench mask in corporation with the field oxide is a step of applying a capacitive-transistor trench mask in an active area isolated as an enclosed area by the field oxide and a gate in the active area. The capacitive-transistor trench mask is employed to define a single edge of the trench capacitor while remaining edges of the trench capacitor are in self-alignment with the field oxide and the gate. The etching edges for the remaining edges are inherently defined in the active area by the field oxide and the gate. In a preferred embodiment, the method further includes steps of: (d) removing the capacitive-transistor trench mask after etching the trench capacitor followed by filling the capacitor trench with a layer of polycrystalline silicon overlaying the active area; and (e) applying the capacitive-transistor trench mask again in opposite polarity relative to the step described above to etch the polycrystalline layer to define a contact opening to the trench capacitor.

According to above drawings and descriptions, this invention also discloses a method for manufacturing a DRAM

24

(dynamic random access memory) cell array on a substrate. The method includes the steps of (a) forming logic transistors on the substrate having polysilicon gates covered by an insulation protective layer wherein the insulation protective layer disposed next to a field oxide layer defining open areas therein-between; and (b) forming trench capacitors for the memory cells by etching the open areas with edges of the trenches defined by the insulation protective layer and the field oxide layer. In a preferred embodiment, the step of forming logic transistors on the substrate having polysilicon gates comprising a step of forming word-line (WL) select transistors each having a WL-transistor gate padded with a WL-select gate-oxide layer having a thickness substantially the same as a gate oxide layer padded under the polysilicon gates of the logic transistors. In another preferred embodiment, the method further includes a step of (c) connecting an error code checking (ECC) and correction means to the memory cell array for checking and correcting substantially all memory read errors within a threshold error-detection-and-correction time. In another preferred embodiment, the method further includes a step of (e) forming a diffusion layer surrounding the trenches having a same conductivity type as a drain of the logic transistors. In another preferred embodiment, the method further includes a step of (f) forming logic transistors on the substrate having polysilicon gates covered by an insulation protective layer; (f) connecting the gate of a plurality of the logic transistors to a ground voltage thus defining a plurality of isolation transistors each separating two adjacent logic transistors wherein the insulation protective layer of the isolation transistors and the adjacent logic transistors defining open areas therein-between; and (g) forming trench capacitors for the memory cells by etching the open areas with edges of the trenches defined by the insulation protective layer of the isolation transistors and the adjacent logic transistors.

An embedded technology of the present invention uses high performance transistor to support both logic circuits and memory circuits. The circuit performance is high, and the manufacture procedures are simple. However, the leakage current caused by the word line transistor is higher than that of prior art word line transistor. Since the thin gate device can not tolerate high voltage operation, we can not use word line boost method to increase storage charge. It is therefore necessary to provide novel design methods to improve the tolerance in leakage current and storage charge loss. U.S. Pat. No. 5,748,547 disclosed methods that can improve signal-to-noise ratio of DRAM array without increasing device area. Using the method, memory devices can be functional without using boosted word line voltages. The same patent disclosed novel self-refresh mechanism that is invisible to external users while using much less power. Using the self-refresh mechanism to increase refresh frequency internally, we can tolerate higher memory leakage current without violating existing memory specifications. Another important method is to use the error-correction-code (ECC) protection to improve the tolerance in non-ideal memory properties.

FIG. 20(a) shows a typical distribution for the refresh time required by the memory cells in a large memory device. For a prior art memory device, the refresh time of the worst bit, i.e., (Tmin), determines its refresh time, among millions of memory cells in the memory device. This worst bit refresh time (Tmin) is typically many orders of magnitudes shorter than the average refresh time (Tav), because the worst bit is always caused by defective structures in the memory cell. FIG. 20(b) shows the simplified block diagram of a memory device equipped with ECC protection circuits. During a



US 6,687,148 B2

25

memory write operation, the input data is processed by a ECC parity tree (2005) to calculate ECC parity data. The input data is stored into a normal data memory array (2001) while the ECC parity data is stored into a parity data array (2003). During a read operation, stored data as well as ECC parity data are read from the memory arrays (2001, 2003) and sent to the ECC parity tree (2005). In case there are corruption data, an ECC correction logic (2007) will find out the problem and correct the error so that the output data will be correct. The ECC correction mechanism is known to the art, but it has not been used on low-cost DRAM because it will require more area. The present invention use ECC protection as a method to improve the tolerance in memory cell leakage current. When a memory device is equipped with an ECC circuit, it will correct most single-bit errors. As a result, the refresh time of the memory device is no longer dependent on the worst bit in the memory. Instead, the device will be function until the errors are more than what the ECC mechanism can correct. The refresh time (T<sub>ecc</sub>) is therefore higher than T<sub>min</sub> as shown in FIG. 20(a).

Base on the above novel design methods, practical memory devices using high performance logic transistor in DRAM memory cells have been manufactured successfully.

Although the present invention has been described in terms of the presently preferred embodiment, it is to be understood that such disclosure is not to be interpreted as limiting. Various alternations and modifications will no doubt become apparent to those skilled in the art after reading the above disclosure. Accordingly, it is intended that the appended claims be interpreted as covering all alternations and modifications as fall within the true spirit and scope of the invention.

I claim:

1. A method for manufacturing a DRAM (dynamic random access memory) cell array which includes a plurality of memory cells, each having a select-transistor comprising:

forming a select-transistor-gate for said select-transistor wherein said select-transistor-gate having substantially a same thickness as a typical transistor of a logic circuit; and

applying implant processes in forming said select-transistor wherein said select-transistor having substantially a same threshold voltage as said typical transistor of a logic circuit.

2. The method for manufacturing said memory cell array of claim 1 further comprising:

applying a capacitive-transistor trench mask for etching a plurality of trench capacitors for said memory cell array.

3. The method for manufacturing said memory cell array of claim 2 wherein

said step of applying a capacitive-transistor trench mask is a step of applying a capacitive-transistor trench mask in an active area isolated by a field oxide wherein said capacitive-transistor trench mask cooperating with said filed oxide for etching said trench in self-alignment in said active area with etching edges defined by said field oxide.

4. The manufacturing said memory cell array of claim 2 wherein:

said step of applying a capacitive-transistor trench mask in corporation with said field oxide is a step of applying a capacitive-transistor trench mask in an active area isolated by said field oxide as an enclosed area wherein said capacitive-transistor trench mask is employed to define a single edge of said trench capacitor while

26

remaining edges of said trench capacitor are in self-alignment with said field oxide wherein said etching edges for said remaining edges are inherently defined in said active area by said filed oxide.

5. The method for manufacturing said memory cell array of claim 2 wherein:

said step of applying a capacitive-transistor trench mask in corporation with said field oxide is a step of applying a capacitive-transistor trench mask in an active area isolated as an enclosed area by said filed oxide and a gate in said active area wherein said capacitive-transistor trench mask is employed to define a single edge of said trench capacitor while remaining edges of said trench capacitor are in self-alignment with said field oxide and said gate wherein said etching edges for said remaining edges are inherently defined in said active area by said field oxide and said gate.

6. The method manufacturing said memory cell array of claim 2 further comprising:

removing said capacitive-transistor trench mask after etching said trench capacitor followed by filling said capacitor trench with a layer of polycrystalline silicon overlaying said active area; and

applying said capacitive-transistor trench mask again in opposite polarity relative to said step in claim 2 to etch said polycrystalline layer to define a contact opening to said trench capacitor.

7. The method manufacturing said memory cell array of claim 1 wherein:

said method further includes a step of manufacturing a DRAM (dynamic random access memory) cell array.

8. The method manufacturing said memory cell array of claim 1 wherein:

said method further includes a step of manufacturing a SRAM (static random access memory) cell array.

9. The method manufacturing said memory cell array of claim 1 wherein:

said method further includes a step of manufacturing a EPROM (erasable programmable read only memory) cell array.

10. The method manufacturing said memory cell array of claim 1 wherein:

said method further includes a step of manufacturing a CAM (content access memory) cell array.

11. The method manufacturing said memory cell array of claim 1 wherein:

said method further includes a step of manufacturing a MRAM (magnetic random access memory) cell array.

12. A method for manufacturing a memory cell array on a substrate comprising:

forming transistors on said substrate wherein each transistor functioning as select transistor for a memory cell of said memory cell array wherein each select transistor having a gate covered by an insulation protective layer wherein said insulation protective layer disposed next to a field oxide layer defining open areas therein-between;

forming trench capacitors for said memory cells by etching said open areas with edges of said trenches defined by said insulation protective layer and said field oxide layer.

13. The method for manufacturing said memory cell array of claim 12 wherein:

said step of forming select transistors on said substrate each having said gate comprising a step of forming



US 6,687,148 B2

27

word-line (WL) select transistors each having a WL-transistor gate padded with a WL-select gate-oxide layer having a thickness substantially the same as a gate oxide layer padded under said gates of said select transistors.

14. The method for manufacturing said memory cell array of claim 13 further comprising:

connecting an error code checking (FCC) and correction means to said memory cell array for checking and correcting substantially all memory read errors within a threshold error-detection-and-correction time.

15. The method for manufacturing said memory cell array of claim 12 further comprising:

forming a diffusion layer surrounding said trenches having a same conductivity type as a drain of said select transistors.

16. The method manufacturing said memory cell array of claim 12 wherein:

said method further includes a step of manufacturing a DRAM (dynamic random access memory) cell array.

17. The method manufacturing said memory cell array of claim 12 wherein:

said method further includes a step of manufacturing a SRAM (static random access memory) cell array.

18. The method manufacturing said memory cell array of claim 12 wherein:

said method further includes a step of manufacturing a EPROM (erasable programmable read only memory) cell array.

19. The method manufacturing said memory cell array of claim 12 wherein:

said method further includes a step of manufacturing a CAM (content access memory) cell array.

20. The method manufacturing said memory cell array of claim 12 wherein:

said method further includes a step of manufacturing a MRAM (magnetic random access memory) cell array.

21. A method for manufacturing a memory cell array on a substrate comprising:

forming a plurality of select transistors on said substrate having polysilicon gates covered by an insulation protective layer;

connecting said gate of a plurality of said logic transistors to a ground voltage thus defining a plurality of isolation transistors each separating two adjacent select transistors wherein said insulation protective layer of said isolation transistors and said adjacent logic transistors defining open areas therein-between;

forming trench capacitors for said memory cells by etching said open areas with edges of said trenches defined

28

by said insulation protective layer of said isolation transistors and said adjacent logic transistors.

22. A memory cell array supported on a substrate comprising:

a plurality of memory cells each having a select-transistor wherein each of said select-transistor having a select-transistor-gate;

said select-transistor-gate having substantially a same thickness as a typical transistor of a logic circuit; and said select-transistor for each of said memory cells having a select-transistor threshold voltage wherein said select-transistor threshold voltage is substantially the same as typical transistor of a logic circuit.

23. The memory cell array of claim 22 wherein:

each of said memory cells further having a trench capacitor.

24. The memory cell array of claim 23 further comprising:

an active area isolated and defined by edges of a field oxide layer disposed on said substrate wherein each of said trench capacitors disposed in said active area and in self-alignment with said edges of said field oxide layer.

25. The memory cell array of claim 23 further comprising:

an active area isolated and defined by edges of a field oxide layer disposed on said substrate wherein each of said trench capacitors disposed in said active area and in self-alignment with said edges of said field oxide layer and edges of said select-transistor gate.

26. The memory cell array of claim 22 further comprising:

an error code checking (ECC) and correction means connected to said memory cell array for checking and correcting substantially all memory read errors within a threshold error-detection-and-correction time.

27. The memory cell array of claim 22 further comprising: a plurality of DRAM (dynamic random access memory) cells.

28. The memory cell array of claim 22 further comprising: a plurality of SRAM (static random access memory) cells.

29. The memory cell array of claim 22 further comprising: a plurality of EPROM (erasable programmable read only memory) cells.

30. The memory cell array of claim 22 further comprising: a plurality of CAM (content access memory) cells.

31. The memory cell array of claim 22 further comprising: a plurality of MRAM (magnetic random access memory) cells.

\* \* \* \* \*

# EXHIBIT C

## CONFIDENTIAL DISCLOSURE AGREEMENT

### (MUTUAL EXCHANGE)

Taiwan Semiconductor Manufacturing Company with a principal place of business as 121, Park Avenue III, Science Based Industrial Park, Hsin-Chu, Taiwan R.O.C. and  
TELESIS INNOVATION INC., with a principal place of business at

991 AMARILLO AVENUE, PALO ALTO, CA 94303

mutually agree that certain confidential information of a party hereto, which if furnished by that party hereunder in written or other tangible form is clearly marked as being confidential, or if orally or visually furnished, is identified as being confidential in a writing submitted to the receiving party within thirty (30) days after such oral or visual disclosure shall be considered by the receiving party to be the Confidential Information of the furnishing party.

Each party agrees to maintain the Confidential Information of the other party received hereunder in confidence utilizing the same degree of care the receiving party uses to protect its own confidential information of a similar nature and to not disclose such information to any third party or to employees of the receiving party without a need to know.

Each party agrees to fully comply with the United States Export Control Regulations, assuring the other party that, unless prior authorization is obtained from the United States Office of Export Administration, the receiving party does not intend to and shall not knowingly export or re-export, directly or indirectly, any Confidential Information received hereunder in contradiction of current Export Administration Regulations published by the United States Department of Commerce.

This Agreement shall impose no obligation upon the receiving party with respect to any Confidential Information of the furnishing party which (i) is now or which subsequently becomes generally known or available; (ii) is known to the receiving party at the time of receipt of same from the furnishing party; (iii) is provided by the furnishing party to a third party without restriction on disclosure; (iv) is subsequently rightfully provided to the receiving party by a third party without restriction or disclosure; or (v) is independently developed by the receiving party provided the person or persons developing same have not had access to the Confidential Information of the furnishing party.

All written data delivered by one party hereto to the other party pursuant to this Agreement shall be and remain the property of the furnishing party, and all such written data, and any copies thereof, shall be promptly returned to the furnishing party upon written request, or destroyed at the furnishing party's option.

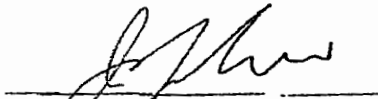
UNIRAM TECHNOLOGY INC.  
CONFIDENTIAL

No rights or obligations other than those expressly recited herein are to be implied from this Agreement. No license is hereby granted directly or indirectly under any patent.

Both parties shall be relieved of all obligations hereunder THREE (3) years after  
SEPTEMBER 16, 1996

UNDERSTOOD AND AGREED:

TELESIS INNOVATION INC.  
Company Name

  
Signature

Jeng-Tye Shan  
Type or Print Name

President  
Title

September 16, 1996  
Date

TAIWAN SEMICONDUCTOR  
MANUFACTURING CO.

  
Signature

John Luke for Quincy Lin  
Type or Print Name

President - TSMC, USA  
Title

SEPTEMBER 16, 1996  
Date

UNIRAM TECHNOLOGY INC.  
CONFIDENTIAL

# EXHIBIT D

## NONDISCLOSURE AGREEMENT

**Taiwan Semiconductor Manufacturing Co., Ltd.**, a company duly incorporated under the laws of the Republic of China, having its principal office located at No. 121, Park Avenue 3, Science Based Industrial Park, Hsinchu, Taiwan ("TSMC"), and **InTempo Technologies, Inc.**, a company duly incorporated under the laws of the State of California, having its principal office located at 991 Amarillo Avenue, Palo Alto, CA 94303, USA ("Company") agree to the following terms and conditions to cover the disclosure of the confidential information described below:

1. The effective date of this Agreement is: October 11, 1999 ("Effective Date").
2. The Confidential Information shall mean information that is disclosed by the disclosing party ("Discloser") to the receiving party ("Recipient") pursuant to this Agreement, including but not limited to, technical, business, financial and marketing information.
3. This Agreement shall be effective and shall govern the disclosure of all Confidential Information between the parties for five (5) years from the Effective Date, unless sooner terminated by either party by giving at least thirty (30) days of prior written notice to the other. The duty of the Recipient to protect the Confidential Information expires five (5) years from the date of disclosure, and shall survive and continue after any termination or expiration of this Agreement.
4. The Recipient shall protect the Confidential Information by using the same degree of care, but no less than a reasonable degree of care, as the Recipient uses to protect its own confidential information of like importance, to prevent unauthorized use, dissemination to any employees of the Recipient without a need to know, communication to any third party or publication of the Confidential Information.
5. The Recipient shall be obligated to treat as the Confidential Information the information that is disclosed by the Discloser either (a) in writing and marked as confidential or similar legend at the time of disclosure; or (b) in any other manner provided it is treated as confidential at the time of disclosure and is summarized and designated as confidential in a written memorandum delivered to the Recipient within thirty (30) days of the disclosure. The Recipient shall only use the Confidential Information disclosed by the Discloser for the purpose(s) specified by the Discloser at the time of disclosure. The Recipient is prohibited from using such Confidential Information for any other purposes.
6. Information shall not be deemed confidential and the Recipient shall have no obligation with respect to any information which (a) is already known to the Recipient without a duty of confidentiality before receipt from the Discloser; (b) is or becomes publicly known through no wrongful act of the Recipient; (c) is rightfully received by the Recipient from a third party without a duty of confidentiality; (d) can be established through the Recipient's written records as having been independently developed by the Recipient; (e) is authorized by the Discloser for release; or (f) is furnished by the Discloser to a third party without a duty of confidentiality. If the Recipient is required to disclose the Confidential Information to a governmental agency or court of law, the Recipient agrees to give the Discloser notice so that the Discloser may contest the disclosure or obtain a protective order.

UNIRAM TECHNOLOGY INC.  
CONFIDENTIAL





7. The Discloser warrants that it has the right to make the disclosures under this Agreement. ALL INFORMATION IS FURNISHED "AS IS". THE DISCLOSER MAKES NO WARRANTY, EXPRESS, IMPLIED OR STATUTORY, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE WITH RESPECT THERETO.
8. All Confidential Information shall be and remain the property of the Discloser. The written Confidential Information, and any copies thereof, shall be promptly returned to the Discloser upon written request, or if destroyed at the Discloser's option, Recipient shall provide a written confirmation for such effect.
9. Neither party acquires any intellectual property rights under this Agreement except the limited right to use the Confidential Information. The Recipient shall cease from using any of the Confidential Information upon the request of the Discloser.
10. Neither party has an obligation under this Agreement to purchase any product or service from the other party or to offer for sale of products using or incorporating the Confidential Information.
11. The parties do not intend to create any agency or partnership relationship between them by this Agreement.
12. The parties agree to take all appropriate measures to comply with the national export control laws, regulations and rules of the Republic of China and the United States of America.
13. Nothing in this Agreement shall be construed as a representation that either party will not independently pursue similar business opportunities, provided that the obligation of this Agreement is not breached.
14. This Agreement is not assignable and constitutes the entire understanding and agreement between the parties as to its subject matter and merges and supersedes all previous communications with respect to their obligations of confidentiality. No addition to or modification of this Agreement shall be binding on either party, unless reduced to writing and signed by each party.
15. This Agreement shall be governed by and construed in accordance with the laws of the Republic of China.

Taiwan Semiconductor Manufacturing  
Co., Ltd.



By: \_\_\_\_\_

Name: Ron Norris

Title: Senior Vice President,  
Worldwide Sales and Marketing

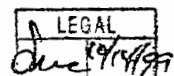
InTempo Technologies, Inc.



Name: J-J Shau

Title: President

UNIRAM TECHNOLOGY INC.  
CONFIDENTIAL





# EXHIBIT E

## NONDISCLOSURE AGREEMENT


Taiwan Semiconductor Manufacturing Co., Ltd., a company duly incorporated under the laws of the Republic of China, having its principal office located at No. 121, Park Avenue 3, Science Based Industrial Park, Hsinchu, Taiwan ( "TSMC" ), and UniRAM Technology, Inc., a company duly incorporated under the laws of California, the United States of America, having its principal office located at 3375 Scott Boulevard, Suite #332, Santa Clara, CA 95054, ( "Company" ) agree to the following terms and conditions for the disclosure of the confidential information described below:

1. This Agreement shall be effective from August 29, 2000, for five (5) years unless sooner terminated by either party by giving a thirty (30) days prior written notice to the other party. Recipient shall keep the Confidential Information in strict confidence for five (5) years from the date of respective disclosure. Recipient acknowledges that its confidentiality obligations under this Agreement shall survive any termination or expiration of this Agreement.
2. "Confidential Information" shall mean information that is disclosed or to be disclosed by the disclosing party ( "Discloser" ) to the receiving party ( "Recipient" ) pursuant to this Agreement, including, but not limited to, technical, business, financial and marketing information solely for the purpose identified herein: providing embedded memory IP for TSMC customers and manufacturing and production of memory component using UniRAM technologies or to be specified by Discloser upon disclosure. Recipient shall only use the Confidential Information for the above purpose(s) and is prohibited from using such Confidential Information for any other purposes.
3. Recipient shall protect the Confidential Information from any misappropriation or unauthorized use by any third party by using the same degree of care, but no less than a reasonable degree of care as the Recipient uses to protect its own confidential information of like importance. Recipient is authorized to disclose the Confidential Information to its employees exclusively on need-to-know basis provided that such employees are subject to valid confidentiality obligations substantiated by written document with terms and conditions substantially similar to those contained in this Agreement. If Recipient desires to disclose the Confidential Information to its customers or third party contractors, the Recipient should first obtain Discloser's prior written consent, then sign with such customers or contractors a valid non-disclosure agreement with terms and conditions substantially similar to those contained in this Agreement. Recipient agrees to be jointly liable for any unauthorized use of the Confidential Information or any type of violation of this Agreement committed by its employees, customers or third party contractors under this Agreement.
4. For purpose of this Agreement, Confidential Information shall be either (a) in writing and marked as confidential or similar legend at the time of disclosure; or (b) in any other manner or media if it is treated as confidential upon disclosure and is designated as confidential in a writing delivered to the Recipient within thirty (30) days after disclosure. Confidential Information shall not include any information which (a) is already known to the Recipient without a duty of confidentiality before the receipt from the Discloser; (b) is or becomes publicly known through no wrongful act of the Recipient; (c) is rightfully received by the Recipient from a third party without a duty of confidentiality; (d) can be sufficiently established, through Recipient's written records, as having been independently developed by the Recipient; (e) is authorized by the Discloser for release; or (f) is furnished by the Discloser to a third party without a duty of confidentiality. If Recipient is required by law or court decrees to disclose the Confidential Information to a governmental agency or court of law, the Recipient agrees to give Discloser proper prior notice so that the Discloser may contest the disclosure or obtain a protective order.

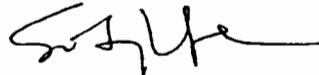


5. Discloser warrants that it has full title, right or license to make the disclosures under this Agreement. ALL INFORMATION IS FURNISHED "AS IS". DISCLOSER DISCLAIMS ANY AND ALL TYPE OF WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE WITH RESPECT THERETO.
6. All Confidential Information shall remain Discloser's property. Upon sooner termination or expiration of this Agreement or subject to Discloser's written request, Recipient shall immediately cease using and return all Confidential Information as well as any copies or duplicates thereof. Alternatively, if Discloser requests the destruction of the Confidential Information, Recipient shall provide Discloser a written affidavit or confirmation signed by its incumbent senior officer to evidence the complete destruction of such Confidential Information.
7. No other licenses or rights, including intellectual property rights, either expressed or implied, are granted under this Agreement except for the Recipient's limited right to use the Confidential Information pursuant to the stipulated purpose(s).
8. Neither party has an obligation under this Agreement to purchase any product or service from the other party or to offer for sale of any products using or incorporating the Confidential Information. Further, the parties do not intend to create any agency or partnership relationship between them by this Agreement.
9. The parties agree to take all appropriate measures to comply with the national export control laws, regulations and rules of the Republic of China and the United States of America.
10. Nothing in this Agreement shall be construed as a representation or inference that either party will not independently pursue similar business opportunities or technology development as long as such activities do not violate this Agreement.
11. This Agreement shall be governed by and construed in accordance with the substantive laws of the Republic of China, without application of its conflict of law rules. Recipient acknowledges that any breach of this agreement could cause irreparable harm and significant injury which monetary damages may be inadequate to remedy. Accordingly, Recipient agrees that Discloser is entitled to seek proper injunctive or equitable relief in any court of competent jurisdiction in addition to any other remedies by operation of laws or in equity.
12. This Agreement is not assignable and shall constitute the entire understanding and agreement between the parties as to its subject matter and merges and supersedes all previous communications with respect to their obligations of confidentiality. No addition to or modification of this Agreement shall be binding on either party, unless reduced to writing and signed by each party.

Taiwan Semiconductor Manufacturing  
Co., Ltd.

By:   
Name: Ron Norris  
Title: Senior Vice President  
Worldwide Marketing & Sales  
Date: 08-31-00'

UniRAM Technology, Inc.

By:   
Name: Sidney Yen  
Title: Senior Director of Operations  
Date: August 29, 2000